Inference in Nonparametric Series Estimation with Specification Searches for the Number of Series Terms

Byunghoon Kang*
Department of Economics, Lancaster University

February 21

Abstract

Nonparametric series regression often involves specification search over the tuning parameter, i.e., evaluating estimates and confidence intervals with a different number of series terms. This paper develops pointwise and uniform inferences for conditional mean functions in nonparametric series estimations that are uniform in the number of series terms. As a result, this paper constructs confidence intervals and confidence bands with possibly data-dependent series terms that have valid asymptotic coverage probabilities. This paper also considers a partially linear model setup and develops inference methods for the parametric part uniform in the number of series terms. The finite sample performance of the proposed methods is investigated in various simulation setups as well as in an illustrative example, i.e., the nonparametric estimation of the wage elasticity of the expected labor supply from Blomquist and Newey (2002).

Keywords: Nonparametric series regression, Pointwise confidence interval, Smoothing parameter choice, Specification search, Undersmoothing, Uniform confidence bands.
JEL classification: C12, C14.

1 Introduction

We consider the following nonparametric regression model

$$y_i = g_0(x_i) + \varepsilon_i, \qquad E(\varepsilon_i|x_i) = 0 \tag{1.1}$$

where $\{y_i, x_i\}_{i=1}^n$ is i.i.d., y_i is a scalar response variable, $x_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a vector of covariates, and $g_0(x) = E(y_i|x_i = x)$ is the conditional mean function. The theory of estimation and inference

*I thank the editor Peter Phillips, the co-editor Iván Fernández-Val, and the two anonymous referees for thoughtful comments that significantly improved this paper. I am also grateful to Bruce Hansen, Jack Porter, Xiaoxia Shi and Joachim Freyberger for useful comments and discussions, and thanks to Michal Kolesár, Denis Chetverikov, Yixiao Sun, Andres Santos, Patrik Guggenberger, Federico Bugni, Joris Pinkse, Liangjun Su, Myung Hwan Seo, and Áureo de Paula for helpful conversations and criticism. This paper is a revised version of the first chapter in my Ph.D. thesis at UW-Madison and previously titled "Inference in Nonparametric Series Estimation with Data-Dependent Undersmoothing". I acknowledge support by the Kwanjeong Educational Foundation Graduate Research Fellowship and Leon Mears Dissertation Fellowship from UW-Madison. All errors are my own. Email: b.kang1@lancaster.ac.uk, Homepage: https://sites.google.com/site/davidbhkang

is well developed for nonparametric series (sieve) methods in a large body of econometrics and statistics literature. Series estimators have also received attention in applied economics because they have many appealing features, e.g., they can easily impose shape restrictions such as additive separability and monotonicity. Once the basis function is chosen (e.g., polynomial or regression spline series of fixed order), implementation requires a choice of the number of series terms $K = K_n$ where K denotes the order of the polynomials or the number of knots in the splines. However, this often involves some ad-hoc specification searches over $K \in \mathcal{K}_n$. For example, when $x_i \in \mathbb{R}^{d_x}$ is vector valued, researchers often evaluate the different numbers of terms in each dimension separately and construct a set of bases with different powers and cross-products of covariates. Although specification search seems necessary in some cases, it may lead to misleading inference without considering the first-step specification search or series term selection.[1]

Existing theory for the asymptotic normality of t-statistics and valid inference imposes a so-called undersmoothing (i.e., overfitting) condition that is a faster rate of K than the mean-squared error (MSE) optimal convergence rates, and many papers in the literature typically suggest rule-of-thumb rules that give the desired level of undersmoothing. Among many others, Newey (2013) suggested increasing K until the standard errors are large relative to small changes in objects of interest. Newey, Powell, and Vella (1999) suggested using more terms than that chosen by cross-validation. Horowitz and Lee (2012) suggested increasing K until the integrated variance suddenly increases and then adding additional terms.

In this paper, we formally justify these rule-of-thumb methods or "plug-in" methods with undersmoothed $\widehat{K}$ for valid inference in nonparametric series regression. Specifically, we provide pointwise inference for $g_0(x)$ with possibly data-dependent (undersmoothed) $\widehat{K} \in \mathcal{K}_n$, i.e., constructing $100(1-\alpha)\%$ confidence interval (CI),

$$\liminf_{n\to\infty} P(g_0(x) \in [\widehat{g}_n(\widehat{K}, x) \pm \widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(\widehat{K}, x)/n}]) \geq 1-\alpha, \tag{1.2}$$

with an estimator $\widehat{g}_n(K, x)$, variance $\widehat{V}_n(K, x)$ using K series terms, and critical values $\widehat{c}_{1-\alpha}(x)$ from the supremum of the t-statistics. For this result, we first develop a uniform distributional approximation theory of the absolute value of the supremum of the t-statistics over different series terms to construct asymptotically valid confidence intervals, which are uniform in $K \in \mathcal{K}_n$,

$$P(g_0(x) \in [\widehat{g}_n(K, x) \pm \widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(K, x)/n}], \quad K \in \mathcal{K}_n) = 1-\alpha+o(1). \tag{1.3}$$

The critical values $\widehat{c}_{1-\alpha}(x)$ can be easily implemented using simple simulation or weighted bootstrap methods.

Furthermore, this paper develops the construction of confidence bands for $g_0(x)$ with asymp-

[1] As a referee noted, the bias and MSE of the series estimator depend on not only K but also the specific bases or sieve spaces, e.g., the order of the splines. In this paper, we fix the basis function, and we do not allow searching over the specific bases or sieve spaces.

totically uniform (in $K \in \mathcal{K}_n$) coverage with critical values $\widehat{c}_{1-\alpha}$ chosen to satisfy

$$P(g_0(x) \in [\widehat{g}_n(K,x) \pm \widehat{c}_{1-\alpha}\sqrt{\widehat{V}_n(K,x)/n}], \quad K \in \mathcal{K}_n, x \in \mathcal{X}) = 1 - \alpha + o(1). \tag{1.4}$$

Analogous to the pointwise inference in (1.2), we can show the validity of confidence bands with the data-dependent $\widehat{K}$. Even in pointwise inference, deriving a uniform asymptotic distribution theory for all sequences of t-statistics over $K \in \mathcal{K}$ may not be possible unless $p = |\mathcal{K}_n|$ is finite. Allowing $p \to \infty$ as $n \to \infty$, results in this paper build on coupling inequalities for the supremum of the empirical process developed by Chernozhukov, Chetverikov, and Kato (2014a, 2016) combined with anti-concentration inequality in Chernozhukov, Chetverikov, and Kato (2014b).

We also provide inference methods in a partially linear model setup focusing on the common parametric part. Unlike the nonparametric object of interest that has a slower convergence rate than $n^{1/2}$ (e.g., regression function or regression derivative), the t-statistics for the parametric object of interest are asymptotically equivalent for all sequences of K under standard rate conditions $K/n \to 0$ as $n \to \infty$. To account for the dependency of the t-statistics with the different sequences of K in this setup, we consider a faster rate of K that grows as fast as the sample size n, as in Cattaneo, Jansson, and Newey (2018a, 2018b), and develop an asymptotic distribution of the t-statistics over $K \in \mathcal{K}_n$. Then, we discuss methods to construct confidence intervals that are similar to the nonparametric regression setup and provide uniform (in $K \in \mathcal{K}_n$) coverage properties.

We investigate finite sample coverage and length properties of the proposed CIs and uniform confidence bands in various simulation setups. As an illustrative example, we revisit nonparametric estimation of labor supply function using the entire individual piecewise-linear budget set as in Blomquist and Newey (2002). Imposing additive separability, which is derived by economic theory, Blomquist and Newey (2002) estimate the conditional mean of labor supply function using series estimation and report wage elasticity of the expected labor supply as well as other welfare measures with various specifications of the different number of series terms.

Several important papers have investigated the asymptotic properties of series (and sieve) estimators, including papers by Andrews (1991a); Eastwood and Gallant (1991); Newey (1997); Chen and Shen (1998); Huang (2003); Chen (2007); Chen and Liao (2014); Chen, Liao, and Sun (2014); Belloni, Chernozhukov, Chetverikov, and Kato (2015); and Chen and Christensen (2015), among many others. This paper extends inference based on the t-statistic under a single sequence of K to the sequences of K over a set $\mathcal{K}_n$ and focuses both on the pointwise and uniform inferences on $g_0(x)$, which is irregular (i.e., slower than a rate of $n^{1/2}$) and a linear functional, under an i.i.d. setup.

The supremum t-statistics have been used as a correction for multiple-testing problems and to construct simultaneous confidence bands, and the importance of multiple-testing problems (data mining or data snooping) has long been noted in various other contexts (see Leamer (1983), White (2000), Romano and Wolf (2005), Hansen (2005)).

There is also a growing literature on data-dependent series term selection and its impact on

estimation and inference in econometrics and statistics. Asymptotic optimality results of cross-validation have been developed, e.g., in papers by Li (1987), Andrews (1991b), and Hansen (2015). Horowitz (2014) develops data-driven methods for choosing the sieve dimension in the nonparametric instrumental variables (NPIV) estimation such that resulting NPIV estimators attain the optimal sup-norm or L^2 norm rates adaptive to the unknown smoothness of $g_0(x)$. Although we do not pursue adaptive inference in this paper, there is also a large statistical literature on adaptive inference. For example, Giné and Nickl (2010), Chernozhukov, Chetverikov, and Kato (2014b) construct adaptive confidence bands in the density estimation problem (see Giné and Nickl (2015, Section 8) for comprehensive lists of references). However, once data-driven choice is obtained for adaptive estimation (e.g., Lepski (1990)-type procedures), one still requires an undersmoothing condition for inference to eliminate asymptotic bias terms (see Theorem 1 of Giné and Nickl (2010)), and this may result in similar specification search issues when choosing sufficiently "large" K in practice.

We can, in principle, consider kernel-based estimation where several data-dependent bandwidth selections or explicit bias corrections have been proposed.[2] However, there exist many examples estimating $g_0(x)$ using (global) series estimation and imposing shape constraints easily (such as additive separability to reduce dimensionality) that are also interested in both pointwise and uniform inference. Given the issues of specification search, our paper is closely related to a recent paper by Armstrong and Kolesár (2018) which considers a bandwidth snooping adjustment for kernel-based inference.

Unlike kernel-based methods, little is known about the statistical properties of data-dependent selection rules and explicit bias formulas for general series estimation. Zhou, Shen, and Wolfe (1998) and Huang (2003) are two of the few exceptions. A recent paper, Cattaneo, Farrell, and Feng (2019), develops novel explicit asymptotic bias/integrated mean squared error (IMSE) formulas and asymptotic theory of the bias-correction methods for general partitioning-based series estimators. The results in Cattaneo, Farrel, and Feng (2019) can be used as an alternative to the undersmoothing approach to avoid specification search issues.

The remainder of the paper is organized as follows. Section 2 introduces the basic nonparametric series regression setup and the candidate set $\mathcal{K}_n$. Section 3 provides the pointwise inference, and Section 4 provides uniform inference in $x \in \mathcal{X}$. Section 5 extends our inference methods to the partially linear model setup. Section 6 summarizes Monte Carlo experiments in various setups, and Section 7 illustrates an empirical example as in Blomquist and Newey (2002). Then, Section 8 concludes the paper. Appendix A includes the main proofs, and Appendix B includes figures and tables. Additional supporting lemmas and simulation results are provided in the Online Supplementary Material available at Cambridge Journals Online (journals.cambridge.org/ect).

[2]See Härdle and Linton (1994), Li and Racine (2007) for references. See also Hall and Horowitz (2013), Calonico, Cattaneo, and Farrell (2018), Schennach (2015) and references therein for various recent works on related bias issues and inference for kernel estimators.

1.1 Notation

$||A||$ denotes the spectral norm, which equals the largest singular value of a matrix A, and $\lambda_{min}(A)$, $\lambda_{max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix A, respectively. $o_p(\cdot)$ and $O_p(\cdot)$ denote the usual stochastic order symbols, $\xrightarrow{d}$ denotes convergence in distribution, and $\Rightarrow$ denotes weak convergence. Let $a \wedge b = \min\{a, b\}, a \vee b = \max\{a, b\}$ and denote $\lfloor a \rfloor$ as the largest integer less than the real number a. For two sequences of positive real numbers a_n and b_n, $a_n \lesssim b_n$ denotes $a_n \leq c b_n$ for all n sufficiently large with some constant $c > 0$ that is independent of n. $a_n \asymp b_n$ denotes $a_n \lesssim b_n$ and $b_n \lesssim a_n$. Furthermore, $a_n \lesssim_P b_n$ denotes $a_n = O_p(b_n)$. For a given random variable $\{X_i\}$ and $1 \leq p < \infty$, $L^p(X)$ is the space of all L^p-norm bounded functions with $||f||_{L^p} = [E||f(X_i)||^p]^{1/p}$, $\ell^\infty(X)$ denotes the space of all bounded functions under the sup-norm, and $||f||_\infty = \sup_{x \in \mathcal{X}} |f(x)|$ for the bounded real-valued functions f on the support $\mathcal{X}$.

2 Setup

We introduce the nonparametric series regression setup in the model (1.1). Given a random sample $\{y_i, x_i\}_{i=1}^n$, we are interested in inference on the conditional mean $g_0(x) = E(y_i|x_i = x)$ at a particular point $x \in \mathcal{X} \subset \mathbb{R}^{d_x}$ or uniform in $x \in \mathcal{X}$.

Let $\widehat{g}_n(K, x)$ be an estimator of $g_0(x)$ using $K = K_n \geq 1$ series terms $P(K, x) = (p_1(x), \cdots, p_K(x))'$, which is a vector of basis functions that can change with n. Standard examples for the basis functions are power series, Fourier series, orthogonal polynomials, splines and wavelets. The series estimator is then obtained by the least square (LS) estimation of y_i on regressors $P(K, x_i)$

$$\widehat{g}_n(K, x) = P(K, x)'\widehat{\beta}_K, \qquad \widehat{\beta}_K = (P^{K\prime} P^K)^{-1} P^{K\prime} Y \tag{2.1}$$

where $P^K = [P_{K1}, \cdots, P_{Kn}]', P_{Ki} \equiv P(K, x_i) = (p_1(x_i), p_2(x_i), \cdots, p_K(x_i))', Y = (y_1, \cdots y_n)'$. Define the least square residuals as $\widehat{\varepsilon}_{Ki} = y_i - P_{Ki}'\widehat{\beta}_K$,

$$\begin{aligned} &\widehat{V}_n(K, x) = P(K, x)'\widehat{Q}_K^{-1}\widehat{\Omega}_K\widehat{Q}_K^{-1}P(K, x), \\ &\widehat{Q}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}', \quad \widehat{\Omega}_K = \frac{1}{n}\sum_{i=1}^n P_{Ki}P_{Ki}'\widehat{\varepsilon}_{Ki}^2, \end{aligned} \tag{2.2}$$

and consider the t-statistic

$$\widehat{T}_n(K, x) \equiv \frac{\sqrt{n}(\widehat{g}_n(K, x) - g_0(x))}{\widehat{V}_n(K, x)^{1/2}}. \tag{2.3}$$

Under standard regularity conditions (discussed in the next section), the t-statistic can be decomposed as follows:

$$\widehat{T}_n(K, x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{P(K, x)'Q_K^{-1}P_{Ki}\varepsilon_i}{\widehat{V}_n(K, x)^{1/2}} - \frac{r_n(K, x)}{\sqrt{\widehat{V}_n(K, x)/n}} + o_p(1) \tag{2.4}$$

where $Q_K = E(P_{Ki}P_{Ki}')$, $r_n(K,x) = g_0(x) - P(K,x)'\beta_K$, and $\beta_K \equiv E[P_{Ki}P_{Ki}'])^{-1}E[P_{Ki}y_i]$ is the best linear L^2 projection coefficient. The first term in the decomposition (2.4) converges to a standard normal distribution for the deterministic sequence $K \to \infty$ as $n \to \infty$, and the second term does not necessarily converge to 0 due to approximation errors $r_n(K,x)$. The second term can be ignored with an undersmoothing assumption, and the asymptotic distribution of the t-statistic, $\widehat{T}_n(K,x) \xrightarrow{d} N(0,1)$, is well known in the literature (see, for examples, Andrews (1991a), Newey (1997), Belloni et al. (2015), and Chen and Christensen (2015), among many others). Then, the $100(1-\alpha)\%$ confidence interval for $g_0(x)$ can be easily constructed using the normal critical value $z_{1-\alpha/2}$

$$\left[\widehat{g}_n(K,x) \pm z_{1-\alpha/2}\sqrt{\widehat{V}_n(K,x)/n}\right]. \tag{2.5}$$

However, it is not clear whether the conventional CI using normal critical values (2.5) has a correct coverage probability with a possibly data-dependent $\widehat{K}$ such as cross-validation or IMSE-optimal selection. First, $\widehat{T}_n(\widehat{K},x) \xrightarrow{d} N(0,1)$ may not hold with a random sequence of $\widehat{K}$, even if we assume the asymptotic bias is negligible. Second, it is well known that some data-dependent rules $\widehat{K}$ do not satisfy the undersmoothing rate conditions, which can lead to a large asymptotic bias and coverage distortion of the standard CI. For example, suppose that the researcher uses $\widehat{K} = \widehat{K}_{\texttt{cv}}$ selected by cross-validation; then, $\widehat{K}_{\texttt{cv}}$ is typically too "small" and violates the undersmoothing assumption needed to ensure the asymptotic normality without bias terms and the valid inference.

As discussed in the introduction, the undersmoothing assumption involves possibly ad-hoc methods to choose series terms K over a *candidate set* $\mathcal{K}_n$ for a valid inference, and cross-validation methods naturally involve specification search over a set of the different number of series terms.

The following set assumption on $\mathcal{K}_n$ is constructed to allow a broad range of K such that $\mathcal{K}_n$ can allow (unknown) an optimal MSE rate of K as well as an undersmoothing rate that increases faster than the optimal MSE rate.

Assumption 2.1. *(Set of number of series terms) Assume the candidate set as* $\mathcal{K}_n = \{K_j : 1 \leq j \leq p\}$*, where* $\underline{K} = K_1 \to \infty$ *and* $\overline{K} = K_p \to \infty$ *as* $n \to \infty$*.*

Here, we consider a possibly growing set of the number of series terms, and a similar assumption is used in the literature, for example, in Newey (1994a, 1994b). Suppose $g_0(x)$ belongs to the Hölder space of smoothness $s > 0$, $\Sigma(s,\mathcal{X})$; then, we obtain optimal L^2 convergence rates $O_p(n^{-s/(2s+d_x)})$ with $K \asymp n^{d_x/(d_x+2s)}$. Assumption 2.1 allows having optimal L^2 rates of K in a large set of classes of functions. By setting $\mathcal{K}_n = [\underline{K},\overline{K}] \cap \mathbb{N}$, $\overline{K} \asymp n^{\overline{\phi}}$ and $\underline{K} \asymp n^{\underline{\phi}}$ with $\overline{\phi} = d_x/(d_x+2\underline{s})$, $\underline{\phi} = d_x/(d_x+2\overline{s})$, Assumption 2.1 contains the number of series terms that obtain an optimal L^2 rate of convergence for $g_0(x) \in \bigcup_{s\in S}\Sigma(s,\mathcal{X})$, $S = [\underline{s},\overline{s}]$. A similar assumption is used in the literature on adaptive inference, although we do not pursue this direction in the current paper.

Assumption 2.1 gives flexible choices of K, as we only assume the rates of K, for example, $\overline{K} = Cn^{\overline{\phi}}, \underline{K} = cn^{\underline{\phi}}$, where c and C can be set arbitrarily small or large. We only require rate restrictions uniformly over $K \in \mathcal{K}$ to guarantee the linearization of the t-statistic in (2.4) and the rates of the cardinality $p = |\mathcal{K}_n|$. Since $K \in \mathcal{K}_n$ is a positive integer and $p \leq \overline{K}$, p is growing at a

rate much slower than n under the rate restrictions in Section 3.

Remark 2.1 ($\mathcal{K}_n$ and the largest K)**.** As a referee noted, specification search is often performed over a simple pre-defined set in practice. For example, a researcher may only use quadratic, cubic, or quartic terms in polynomial regression or try only a few different numbers of knots in regression splines to observe how the estimate and standard error change. In the nonparametric estimation of the Mincer equation (Heckman, Lochner, and Todd (2006)), researchers may consider a regression of log wages on experience with polynomials of order $\underline{K} = 1$ (linear) to $\overline{K} = 4$ (quartic).[3]

However, it may not be clear how to define *a priori* $\mathcal{K}_n$ in practice. One must first consider a set of pre-selected models over which to search. As discussed earlier and suggested by many papers in the literature, some formal data-dependent methods to obtain optimal L^2 norm or sup-norm rates, such as cross-validation, can be a useful guideline for $\mathcal{K}_n$. For example, one can consider a reasonable set $\widetilde{\mathcal{K}}_n$ first, choose $\widehat{K}_{\mathtt{cv}} \in \widetilde{\mathcal{K}}_n$ by cross-validation, and then consider $\mathcal{K}_n = [\widehat{K}_{\mathtt{cv}}, c_1\widehat{K}_{\mathtt{cv}}]$ or $[\widehat{K}_{\mathtt{cv}}, \widehat{K}_{\mathtt{cv}}n^{c_2}]$ for some constants $c_1, c_2 > 0$. One can also search $\underline{K}$ and $\overline{K}$ sequentially by calculating changes in cross-validation or standard errors from the initial candidate set. Extending results developed in this paper with data-dependent $\mathcal{K}_n$ are beyond the scope of the paper.

3 Pointwise Inference

In this section, we focus on pointwise inference for $g_0(x)$. The goal of this section is to provide a uniform distributional approximation theory of $\widehat{T}_n(K, x)$ over a set $\mathcal{K}_n$ and provide uniform (in $K \in \mathcal{K}_n$) coverage properties of confidence intervals for $g_0(x)$ in (1.2), (1.3) with the construction of critical values.

From the decomposition of the t-statistic in (2.4), we first consider the (infeasible) test statistic

$$\max_{K\in\mathcal{K}_n} |t_n(K, x)| = \max_{1\le j\le p} |t_n(K_j, x)| \tag{3.1}$$

where $t_n(K, x) = n^{-1/2}\sum_{i=1}^n P(K, x)'Q_K^{-1}P_{Ki}\varepsilon_i/V_n(K, x)^{1/2}$ with the series variance $V_n(K, x) = P(K, x)'Q_K^{-1}\Omega_K Q_K^{-1}P(K, x)$, $\Omega_K = E(P_{Ki}P_{Ki}'\varepsilon_i^2)$. In general, $t_n(K, x), K \in \mathcal{K}_n$ does not have a limiting distribution because it is not asymptotically tight under Assumption 2.1 unless $|\mathcal{K}_n|$ is finite or under the restrictive assumption on $\mathcal{K}_n$.[4] However, we show below that there exists a sequence of random variables $\max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|$ such that $\left|\max_{K\in\mathcal{K}_n}|t_n(K, x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\right| = O_p(a_n)$ for a sequence of constants $a_n \to 0$, where $Z_i = (Z_{i1}, ..., Z_{ip})'$ is a Gaussian random vector in $\mathbb{R}^p$ such that $Z_i \sim N(0, \frac{1}{n}\Sigma_n)$ with (j, l) elements of the variance-covariance matrix

$$\Sigma_n(j, l) = E[t_n(K_j, x)t_n(K_l, x))] = \frac{P(K_j, x)'Q_{K_j}^{-1}\Omega_{K_j,K_l}Q_{K_l}^{-1}P(K_l, x)}{V_n(K_j, x)^{1/2}V_n(K_l, x)^{1/2}}, \tag{3.2}$$

[3] All of our results continue to hold with fixed p; however, it may be preferred to use larger sets $\mathcal{K}_n$ with $p \to \infty$ to give greater flexibility to the candidate models as the sample size n increases.

[4] In an earlier version of the paper, we provide the weak convergence of a series process under the same rates of $K \in \mathcal{K}_n$ and high-level assumptions. This can be viewed as an analogous result in the kernel estimation literature (see Section 2 of Armstrong and Kolesár (2018) and other references therein).

$\Omega_{K_j,K_l} = E(P_{K_ji}P'_{K_li}\varepsilon_i^2)$.

By replacing unknown $\Sigma_n, V_n(K,x)$ with consistent estimators $\widehat{\Sigma}_n, \widehat{V}_n(K,x)$, we show below that we can approximate $\max_{K\in\mathcal{K}_n}|\widehat{T}_n(K,x)|$ by $\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|$ and then obtain critical values by using a simulation-based method to provide valid coverage properties in (1.2) and (1.3). We define $\widehat{c}_{1-\alpha}(x)$ as follows:

$$\begin{aligned}
&\widehat{c}_{1-\alpha}(x) \equiv (1-\alpha) \text{ quantile of } \max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}|, \text{ where } \widehat{Z}_i = (\widehat{Z}_{i1},...,\widehat{Z}_{ip})' \sim N(0,\frac{1}{n}\widehat{\Sigma}_n),\\
&\widehat{\Sigma}_n(j,j) = 1, \quad \widehat{\Sigma}_n(j,l) = \frac{\widehat{V}_n(K_j,K_l,x)}{\widehat{V}_n(K_j,x)^{1/2}\widehat{V}_n(K_l,x)^{1/2}},\\
&\widehat{V}_n(K_j,K_l,x) = P(K_j,x)'\widehat{Q}_{K_j}^{-1}\widehat{\Omega}_{K_j,K_l}\widehat{Q}_{K_l}^{-1}P(K_l,x), \widehat{\Omega}_{K_j,K_l} = \frac{1}{n}\sum_{i=1}^n P_{K_ji}P'_{K_li}\widehat{\varepsilon}_{K_ji}\widehat{\varepsilon}_{K_li}
\end{aligned} \tag{3.3}$$

where $\widehat{\Sigma}_n$ is a consistent estimator of the variance-covariance matrix Σ_n defined in (3.2), $\widehat{V}_n(K,x)$ is the simple plug-in estimator for $V_n(K,x)$ as in (2.2), and $\widehat{\varepsilon}_{Ki} = y_i - P'_{Ki}\widehat{\beta}_K, \forall K \in \mathcal{K}_n$. One can compute $\widehat{c}_{1-\alpha}(x)$ by simulating B (typically $B = 1000$ or 5000) i.i.d. random vectors $\widehat{Z}_i^b \sim N(0, \frac{1}{n}\widehat{\Sigma}_n)$ and by taking a $(1-\alpha)$ sample quantile of $\{\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}^b| : b = 1,\cdots,B\}$. Alternatively, we can use weighted bootstrap methods. See Section 4 for the implementation and the validity of bootstrap procedures in the construction of confidence bands.

To establish our main results, we impose mild regularity conditions uniform in $K \in \mathcal{K}_n$. For each $K \in \mathcal{K}_n$, define $\zeta_K \equiv \sup_{x\in\mathcal{X}}||P(K,x)||$ as the largest normalized length of the regressor vector and $\lambda_K \equiv (\lambda_{min}(Q_K))^{-1/2}$ for $K \times K$ design matrix $Q_K = E(P_{Ki}P'_{Ki})$.

Assumption 3.1. *(Regularity conditions - model)*

(i) $\{y_i, x_i\}_{i=1}^n$ *are i.i.d random variables satisfying the model* (1.1).

(ii) $\max_{K\in\mathcal{K}_n}\lambda_K \lesssim 1$, *and for each* $K \in \mathcal{K}_n$, *as* $K \to \infty$, *there exists* c_K, ℓ_K *such that*

$$\sup_{x\in\mathcal{X}}|r_n(K,x)| \le \ell_K c_K, \quad E[r_n(K,x)^2]^{1/2} \le c_K,$$

where $r_n(K,x) = g_0(x) - P(K,x)'\beta_K, \beta_K = (E[P_{Ki}P'_{Ki}])^{-1}E[P_{Ki}y_i]$.

Assumption 3.2. *(Regularity conditions - pointwise inference)*

(i) $\max_{K\in\mathcal{K}_n}\sqrt{\zeta_K^2\log K\log^2 p/n}(1+\sqrt{K}\ell_K c_K) + \ell_K c_K\log p \to 0$ *as* $n \to \infty$.

(ii) $\sup_{x\in\mathcal{X}}E(|\varepsilon_i|^3|x_i = x) < \infty$, $\inf_{x\in\mathcal{X}}E(\varepsilon_i^2|x_i = x) > 0$, *and either of the following conditions hold: (a)* $\sup_{x\in\mathcal{X}}E[|\varepsilon_i|^q|x_i = x] < \infty$ *for* $q \ge 4$ *or (b) there exists a constant* $C > 0$ *such that* $\sup_{x\in\mathcal{X}}E[\exp(|\varepsilon_i|/C)|X_i = x] \le 2$.

(iii) $\max_{K\in\mathcal{K}_n}|\frac{V_n(K,x)}{\widehat{V}_n(K,x)} - 1| = o_p(1/\log p)$, $\max_{1\le j,l\le p}|\widehat{\Sigma}_n(j,l) - \Sigma_n(j,l)| = o_p(1/\log^2 p)$.

Assumptions 3.1(ii) and 3.2(i) are similar to those imposed in Belloni et al. (2015) and Chen and Christensen (2015), and all the discussions made there also apply here except that we impose rate conditions of K uniformly over $\mathcal{K}_n$. The rate conditions can be replaced by the specific bounds of ζ_K, c_K, ℓ_K with various sieve bases. For example, when $\mathcal{X} = [0,1]^{d_x}$, the probability density of x_i is uniformly bounded above and bounded away from zero, and $g_0(x) \in \Sigma(s, \mathcal{X})$, i.e., the Hölder space of smoothness $s > 0$, then $\lambda_K \lesssim 1$, $\zeta_K \lesssim \sqrt{K}$, $\ell_K c_K \lesssim K^{-(s \wedge s_0)/d_x}$ for regression spline series of order s_0, and Assumption 3.2(i) is satisfied when $\sqrt{\overline{K}(\log^3 \overline{K})/n}(1 + \overline{K}^{1/2}\underline{K}^{-(s\wedge s_0)/d_x}) + \underline{K}^{-(s\wedge s_0)/d_x} \log \overline{K} \to 0$. Other standard regularity conditions in the literature (e.g., Newey (1997) and Chen (2007)) can also be used here, and the rate condition can be improved with different pointwise linearization and approximation bounds in Huang (2003) for splines and Cattaneo et al. (2019) for partitioning-based estimators.

Assumption 3.2(ii) imposes either the bounded polynomial moment conditions or sub-exponential moments of the regression errors. Assumption 3.2(iii) imposes the consistency of variance estimator $\widehat{V}_n(K,x)$ uniformly in $K \in \mathcal{K}_n$, and this holds under mild regularity conditions (see Lemma 5.1 of Belloni et al. (2015) and Lemma 3.1-3.2 of Chen and Christensen (2015)).

Theorem 3.1. *Suppose that Assumptions 2.1, 3.1, and 3.2 hold and that either of the following rate conditions hold depending on the case (a) or (b) in Assumption 3.2(ii): (a)* $(\max_K \zeta_K)^2 \log^5 n \log^3 p/n \vee \max_K \zeta_K \log^{3/4} n \log p/n^{1/2-1/q} \to 0$ *or (b)* $(\max_K \zeta_K)^2 \log^5 n \log^3 p/n \to 0$. *If, in addition, we assume that* $\max_{K\in\mathcal{K}_n} |\frac{\sqrt{n}r_n(K,x)}{V_n(K,x)^{1/2}}| = o(1/\sqrt{\log p})$, *then*

$$\sup_{u\in\mathbb{R}} \big| P(\max_{K\in\mathcal{K}_n} |\widehat{T}_n(K,x)| \le u) - P(\max_{1\le j\le p} \sum_{i=1}^{n} |\widehat{Z}_{ij}| \le u) \big| = o(1), \tag{3.4}$$

and the following coverage property holds

$$P(g_0(x) \in [\widehat{g}_n(K,x) \pm \widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(K,x)/n}], \quad K \in \mathcal{K}_n) = 1 - \alpha + o(1) \tag{3.5}$$

with a critical value $\widehat{c}_{1-\alpha}(x)$ *defined in* (3.3). *Alternatively, if we assume* $|\frac{\sqrt{n}r_n(\widehat{K},x)}{V_n(\widehat{K},x)^{1/2}}| = o(1/\sqrt{\log p})$ *with* $\widehat{K} \in \mathcal{K}_n$, *then the following holds:*

$$\liminf_{n\to\infty} P(g_0(x) \in [\widehat{g}_n(\widehat{K},x) \pm \widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(\widehat{K},x)/n}]) \ge 1 - \alpha. \tag{3.6}$$

Theorem 3.1 provides a uniform coverage property of the confidence interval over $K \in \mathcal{K}_n$ for the regression function $g_0(x)$. Equation (3.6) guarantees the asymptotic coverage of CI for data-dependent $\widehat{K} \in \mathcal{K}_n$ with undersmoothing. Note that standard inference methods in the nonparametric regression setup typically consider a singleton set $\mathcal{K}_n = \{K\}$ with $K \to \infty$ as $n \to \infty$. The rate restriction is mild because it only requires $\overline{K}/n^{1-2/q} \to 0$, up to $\log n$ terms, in case (a) and $\overline{K}/n \to 0$, up to $\log n$ terms, in case (b) when $\zeta_K \lesssim \sqrt{K}$ for splines and wavelet series. Theorem 3.1 builds upon a coupling inequality for maxima of sums of random vectors in

Chernozhukov, Chetverikov, and Kato (2014a) combined with the anti-concentration inequality in Chernozukhov, Chetverikov, and Kato (2014b).

Remark 3.1 (Undersmoothing assumption)**.** Note that (3.5) requires an undersmoothing assumption uniformly over $K \in \mathcal{K}_n$. Without $\max_{K \in \mathcal{K}_n} |\frac{\sqrt{n} r_n(K,x)}{V_n(K,x)^{1/2}}| = o(1)$, coverage in (3.5) can be understood as the uniform confidence intervals for the pseudo-true value $g(K,x) = P(K,x)'\beta_K$, i.e.,

$$P(g(K,x) \in [\widehat{g}_n(K,x) \pm \widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(K,x)/n}], \quad K \in \mathcal{K}_n) = 1 - \alpha + o(1) \tag{3.7}$$

However, a uniform undersmoothing condition is not assumed in (3.6), and it only requires that the chosen $\widehat{K} \in \mathcal{K}_n$ satisfies the undersmoothing condition such that the asymptotic bias is negligible. This allows broader ranges of K in $\mathcal{K}_n$ including an unknown optimal MSE rate. We formally justify rule-of-thumb methods for valid inference suggested in the literature that include an additional number of series terms, a blow up of the numbers after using cross-validation, or some "plug-in" methods for choosing $\widehat{K}$ such as those in Newey, Powell, and Vella (1999), Newey (2013). Here, uniform (in $K \in \mathcal{K}_n$) inference considers uncertainty from specification search and using larger critical values $\widehat{c}_{1-\alpha}(x)$ than the normal critical value $z_{1-\alpha/2}$.

Remark 3.2 (Other functionals)**.** Here, we focus on the leading example with $g_0(x)$ for some fixed point $x \in \mathcal{X}$; however, we can consider other linear functionals $a(g_0(\cdot))$ such as the regression derivatives $a(g_0(x)) = \frac{d}{dx} g_0(x)$. All the results in this paper can be applied to irregular (slower than $n^{1/2}$ rate) linear functionals using estimators $a(\widehat{g}_n(K,x)) = a_K(x)'\widehat{\beta}_K$ and an appropriate transformation of basis $a_K(x) = (a(p_1(x), \cdots, a(p_K(x)))'$ with proper smoothness condition on the functional and continuity conditions on the derivative as in Newey (1997). Although the verification of previous results for regular ($n^{1/2}$ rate) functionals, such as integrals and weighted average derivatives, is beyond the scope of this paper, we examine similar results for the partially linear model setup in Section 5.

4 Uniform Inference

This section provides construction of uniform confidence bands for $g_0(x)$ (uniform in $K \in \mathcal{K}_n$) given in (1.4). We define the following empirical process

$$\widehat{T}_n(K,x) \equiv \frac{\sqrt{n}(\widehat{g}_n(K,x) - g_0(x))}{\widehat{V}_n(K,x)^{1/2}} \tag{4.1}$$

over $\mathcal{K}_n \times \mathcal{X}$, and we show below that the supremum of the empirical process $\sup_{(K,x) \in \mathcal{K}_n \times \mathcal{X}} |\widehat{T}_n(K, x)|$ can be approximated by a sequence of random variables $\sup_{(K,x) \in \mathcal{K}_n \times \mathcal{X}} |Z_n(K,x)|$, where $Z_n(K, x)$ is a tight Gaussian random process in $\ell^\infty(\mathcal{K}_n \times \mathcal{X})$ with zero mean and covariance function

$$E[Z_n(K,x)Z_n(K',x')] = \frac{P(K,x)'Q_K^{-1}\Omega_{K,K'}Q_{K'}^{-1}P(K',x')}{V_n(K,x)^{1/2}V_n(K',x')^{1/2}}. \tag{4.2}$$

Although the Gaussian approximation is an important first step, the covariance function (4.2) is generally difficult to construct for the purpose of uniform inference. Thus, we employ weighted bootstrap methods similar to Belloni et al. (2015) and show the validity of the bootstrap procedure for uniform confidence bands.

Let $e_1, ..., e_n$ be a sequence of i.i.d. standard exponential random variables that are independent of $X^n = \{x_1, ..., x_n\}$. For $(K, x) \in \mathcal{K}_n \times \mathcal{X}$, we define a (centered) weighted bootstrap process

$$\widehat{T}_n^e(K,x) = \frac{\sqrt{n}(\widehat{g}_n^e(K,x) - \widehat{g}_n(K,x))}{\widehat{V}_n(K,x)^{1/2}} \tag{4.3}$$

where $\widehat{g}_n^e(K,x) = P(K,x)'\widehat{\beta}_K^e$, and $\widehat{\beta}_K^e$ is obtained by the following weighted least squares regression

$$\widehat{\beta}_K^e = \underset{\beta \in \mathbb{R}^K}{\arg\min} \sum_{i=1}^n e_i (y_i - P(K,x_i)'\beta)^2. \tag{4.4}$$

Define the critical value

$$\widehat{c}_{1-\alpha} \equiv (1-\alpha) \text{ conditional quantile of } \sup_{K \in \mathcal{K}_n, x \in \mathcal{X}} |\widehat{T}_n^e(K,x)| \text{ given the data } X^n, \tag{4.5}$$

and we consider confidence bands of the form

$$[\widehat{g}_n(K,x) \pm \widehat{c}_{1-\alpha} \sqrt{\widehat{V}_n(K,x)/n}], \quad K \in \mathcal{K}_n, x \in \mathcal{X}. \tag{4.6}$$

To provide the validity of the bootstrap critical values and confidence bands in (4.6), we show below that the conditional distribution of $\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n^e(K,x)|$ is "close" to the distribution of $\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |Z_n(K,x)|$ and that of $\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n(K,x)|$ using coupling inequalities for the supremum of the empirical process and the bootstrap process as in Chernozhukov et al. (2016). Then, similar to Theorem 3.1, this gives bounds on the Kolmogorov distance for the distribution functions of $P(\sup_{K\in\mathcal{K}_n, x\in\mathcal{X}} |\widehat{T}_n(K,x)| \le u)$ and $P(\sup_{K\in\mathcal{K}_n, x\in\mathcal{X}} |\widehat{T}_n^e(K,x)| \le u | X^n)$.

The following assumptions are used to establish the coverage probability of confidence bands uniformly over $K \in \mathcal{K}_n$. Define $\alpha(K,x) \equiv Q_K^{-1/2} P(K,x)/V_n(K,x)^{1/2}$, and

$$\zeta^{L_1} = \max_{K\in\mathcal{K}_n} \sup_{x,x'\in\mathcal{X}, x\neq x'} \frac{||\alpha(K,x) - \alpha(K,x')||}{||x - x'||}, \ \zeta^{L_2} = \sup_{x\in\mathcal{X}} \max_{K,K'\in\mathcal{K}_n : K\neq K'} \frac{||\alpha(K,x) - \alpha(K',x)||}{|K - K'|}.$$

Assumption 4.1. *(Regularity conditions - uniform inference)*

(i) $\sup_x E[|\varepsilon_i|^q | x_i = x] < \infty$ *for* $q \ge 4$ *and* $\inf_{x\in\mathcal{X}} E(\varepsilon_i^2 | x_i = x) > 0$.

(ii) $\max_{K\in\mathcal{K}_n} \sqrt{\frac{\lambda_K^2 \zeta_K^2 \log K \log^4 n}{n}} (n^{1/q} + \ell_K c_K \sqrt{K}) + (\ell_K c_K) \log n \to 0$ *as* $n \to \infty$.

(iii) $\log(\zeta^{L_1} \vee \zeta^{L_2}) \lesssim \log n$, $\max_K \zeta_K^{2q/(q-2)} \log^3 n / n \lesssim 1$, *and* $\max_K \zeta_K \lesssim \log n$.

(iv) $\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\frac{V_n(K,x)}{\widehat{V}_n(K,x)} - 1| = o_p(1/\log^2 n)$.

For uniform inference, we require similar but slightly stronger conditions compared to Assumption 3.2. We also impose mild rate restrictions on ζ^{L_1}, ζ^{L_2} and $\max_{K\in\mathcal{K}_n}\zeta_K$ similar to Chernozhukov et al. (2014a) and Belloni et al. (2015).

Theorem 4.1. *Suppose that Assumptions 2.1, 3.1, and 4.1 hold, and* $(\max_K \zeta_K)\log^{2+1/(2q)} n/n^{1/2-1/q} \to 0$, $(\max_K \zeta_K)^2 \log^7 n/n \to 0$ *as* $n \to \infty$. *If, in addition, we assume that* $\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\frac{\sqrt{n}r_n(K,x)}{V_n(K,x)^{1/2}}| = o(1/\sqrt{\log n})$, *then*

$$P(g_0(x) \in [\widehat{g}_n(K,x) \pm \widehat{c}_{1-\alpha}\sqrt{\widehat{V}_n(K,x)/n}], \quad K \in \mathcal{K}_n, x \in \mathcal{X}) = 1-\alpha+o(1) \tag{4.7}$$

with a critical value $\widehat{c}_{1-\alpha}$ *in* (4.5).

Alternatively, if we assume $\sup_{x\in\mathcal{X}} |\frac{\sqrt{n}r_n(\widehat{K},x)}{V_n(\widehat{K},x)^{1/2}}| = o(1/\sqrt{\log n})$ *with* $\widehat{K} \in \mathcal{K}_n$, *then the following coverage property holds:*

$$\liminf_{n\to\infty} P(g_0(x) \in [\widehat{g}_n(\widehat{K},x) \pm \widehat{c}_{1-\alpha}\sqrt{\widehat{V}_n(\widehat{K},x)/n}], \quad x \in \mathcal{X}) \geq 1-\alpha. \tag{4.8}$$

Theorem 4.1 shows the uniform asymptotic coverage property of the confidence bands defined in (4.6) uniformly over $K \in \mathcal{K}_n$. Furthermore, it shows a confidence band with possibly data-dependent $\widehat{K} \in \mathcal{K}_n$ having an asymptotic coverage of at least $1-\alpha$. The confidence band constructed in (4.8) requires a substantially weaker assumption on the undersmoothing similar to Theorem 3.1.

5 Extension: Partially Linear Model

In this section we provide inference methods for the partially linear model (PLM) setup. For notational simplicity, we use similar notation as defined in the nonparametric regression setup. Suppose we observe random samples $\{y_i, w_i, x_i\}_{i=1}^n$, where y_i is the scalar response variable, $w_i \in \mathcal{W} \subset \mathbb{R}$ is the treatment/policy variable of interest, and $x_i \in \mathcal{X} \subset \mathbb{R}^{d_x}$ is a set of explanatory variables. For simplicity, we shall assume that w_i is a scalar. We consider the model

$$y_i = \theta_0 w_i + g_0(x_i) + \varepsilon_i, \qquad E(\varepsilon_i|w_i, x_i) = 0. \tag{5.1}$$

We are interested in inference on θ_0 after approximating an unknown function $g_0(x)$ by series terms/regressors $p(x_i)$ among a set of potential control variables. Specification searches can be performed for the number of different approximating terms or for the number of covariates in estimating the nonparametric part.

The series estimator $\widehat{\theta}_n(K)$ for θ_0 using the first $K = K_n$ terms is obtained by standard LS estimation of y_i on w_i and $P_{Ki} = P(K, x_i)$ and has the usual "partialling out" formula

$$\widehat{\theta}_n(K) = \left(W' M_K W\right)^{-1} W' M_K Y \tag{5.2}$$

where $W = (w_1, \cdots, w_n)', M_K = I_K - P^K(P^{K\prime}P^K)^{-1}P^{K\prime}, P^K = [P_{K1}, \cdots, P_{Kn}]', Y = (y_1, \cdots, y_n)'$.

The asymptotic normality and valid inference for $\widehat{\theta}_n(K)$ have been developed in the literature.[5] Donald and Newey (1994) derived the asymptotic normality of $\widehat{\theta}_n(K)$ under standard rate conditions $K/n \to 0$. Belloni, Chernozukhov, and Hansen (2014) analyzed asymptotic normality and uniformly valid inference for the post-double-selection estimator even when K is much larger than n (see also Kozbur (2018)). Recent papers by Cattaneo, Jansson, and Newey (2018a, 2018b) provided a valid approximation theory for $\widehat{\theta}_n(K)$ when K grows at the same rate of n.

A different approximation theory using a faster rate of K ($K/n \to c, 0 < c < 1$) than the standard rate conditions ($K/n \to 0$) is particularly useful for our purpose to establish the asymptotic distribution of t-statistics over $K \in \mathcal{K}_n$. From the results in Cattaneo, Jansson, and Newey (2018a), we have the following decomposition:

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_n(K) - \theta_0) &= (\frac{1}{n}W'M_KW)^{-1}\frac{1}{\sqrt{n}}W'M_KY \\ &= \widehat{\Gamma}_n(K)^{-1}(\frac{1}{\sqrt{n}}\sum_i v_iM_{K,ii}\varepsilon_i + \frac{1}{\sqrt{n}}\sum_{i=1}^n\sum_{j=1,j\neq i}^n v_iM_{K,ij}\varepsilon_j) + o_p(1) \qquad (5.3)\end{aligned}$$

where $v_i \equiv w_i - g_{w0}(x_i)$, $g_{w0}(x_i) \equiv E[w_i|x_i]$ and $\widehat{\Gamma}_n(K) = W'M_KW/n$. For any deterministic sequence $K \to \infty$ satisfying standard rate conditions $K/n \to 0$, $\sqrt{n}(\widehat{\theta}_n(K) - \theta_0)$ is asymptotically normal with variance $V = \Gamma^{-1}\Omega\Gamma^{-1}, \Gamma = E[v_iv_i'], \Omega = E[v_iv_i'\varepsilon_i^2]$. Unlike the nonparametric object of interest in the fully nonparametric model, where the variance term increases with K, $\widehat{\theta}_n(K)$ has a parametric ($n^{1/2}$) convergence rate, and $\widehat{\theta}_n(K)$ with all different sequences of K are asymptotically equivalent under $K/n \to 0$.[6] However, under faster rate conditions, $K/n \to c$ for $0 < c < 1$, the second term in (5.3) is not negligible and converges to bounded random variables. Cattaneo, Jansson, and Newey (2018a) apply the central limit theorem of degenerate U-statistics for the second term, similar to the many instrument asymptotics analyzed in Chao, Swanson, Hausman, Newey, and Woutersen (2012). Then, the limiting normal distribution has a larger variance than the standard first-order asymptotic variance, and the adjusted variances generally depend on the number of terms K such that we can provide an asymptotic distribution of the t-statistics with the different sequence of K over $\mathcal{K}_n$.

The following assumption on $\mathcal{K}_n$ is considered, and we impose the regularity conditions that are used in Cattaneo, Jansson, and Newey (2018a, Assumption PLM) uniformly over $K \in \mathcal{K}_n$.

Assumption 5.1. *(Set of finite number of series terms)*

Assume $\mathcal{K}_n = \{\underline{K} \equiv K_1, \cdots, K_m, \cdots, \overline{K} \equiv K_p\}$, *where* $K_m \to \infty, K_m/n \to c_m$ *as* $n \to \infty$ *for all* $m = 1, ..., p$, *constant* c_m *such that* $0 < c_1 < c_2 < \cdots < c_p < 1$, *and fixed* p.

Assumption 5.2. *(Regularity conditions - partially linear model)*

[5] See also Robinson (1988), Linton (1995) and references therein for the results of the kernel estimators.

[6] This is also related to the well-known results of the two-step semiparametric estimation; the asymptotic variance of two-step semiparametric estimators does not depend on the type of the first-step estimator or smoothing parameter sequences under certain conditions (see Newey (1994b)).

(i) $\{y_i, w_i, x_i\}_{i=1}^n$ *are i.i.d random variables satisfying the model* (5.1).

(ii) *There exists constants* $0 < c \leq C < \infty$ *such that* $E[\varepsilon_i^2|w_i, x_i] \geq c$ *and* $E[v_i^2|x_i] \geq c$, $E[\varepsilon_i^4|w_i, x_i] \leq C$ *and* $E[v_i^4|x_i] \leq C$.

(iii) $rank(P_K) = K$ *(a.s.) and* $M_{K,ii} \geq C$ *for* $C > 0$ *for all* $K \in \mathcal{K}_n$.

(iv) *For each* $K \in \mathcal{K}_n$, *there exists some* γ_g, γ_{g_w},

$$\min_{\eta_g} E[(g_0(x_i) - \eta_g' P_{Ki})^2] = O(K^{-2\gamma_g}), \quad \min_{\eta_{g_w}} E[(g_{w0}(x_i) - \eta_{g_w}' P_{Ki})^2] = O(K^{-2\gamma_{g_w}}).$$

Assumption 5.2 does not require $K/n \to 0$ which is required to obtain asymptotic normality in the literature (e.g., Donald and Newey (1994)). Similar to Assumption 3.2(iii) in the nonparametric setup, Assumption 5.2(iv) holds for the polynomials and spline basis. For example, 5.2(iv) holds with $\gamma_g = s_g/d_x, \gamma_{g_w} = s_w/d_x$ when $\mathcal{X}$ is compact and when the unknown functions $g_0(x)$ and $g_{w0}(x)$ have s_g and s_w continuous derivates, respectively.

Under Assumptions 5.1, 5.2 and undersmoothing condition ($n\overline{K}^{-2(\gamma_g+\gamma_{g_w})} \to 0$), we have a joint asymptotic distribution of the t-statistics $T_n(K, \theta) = \sqrt{n} V_n(K)^{-1/2}(\widehat{\theta}_n(K) - \theta_0)$ over $K \in \mathcal{K}_n$:

$$(T_n(K_1, \theta_0), \cdots, T_n(K_p, \theta_0))' \xrightarrow{d} Z_\Sigma = (Z_1, \cdots Z_p)' \sim N(0, \Sigma)$$

where

$$\begin{aligned} &V_n(K) = \Gamma_n(K)^{-1}\Omega_n(K)\Gamma_n(K)^{-1}, \\ &\Gamma_n(K) = \frac{1}{n}\sum_{i=1}^n M_{K,ii} E[v_i^2|x_i], \ \Omega_n(K) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n M_{K,ij}^2 E[v_i^2\varepsilon_j^2|x_i, x_j], \end{aligned}$$

and the variance-covariance matrix Σ with (l, l') element

$$\begin{aligned} &\Sigma(l, l') \equiv \lim_{n\to\infty} \frac{V_n(K_l, K_{l'})}{V_n(K_l)^{1/2}(K_{l'})^{1/2}}, \quad V_n(K_l, K_{l'}) = \Gamma_n(K_l)^{-1}\Omega_n(K_l, K_{l'})\Gamma_n(K_{l'})^{-1} \\ &\Omega_n(K_l, K_{l'}) = \frac{1}{n}\sum_{i=1}^n \sum_{j=1}^n M_{K_l,ij} M_{K_{l'},ij} E[v_i^2\varepsilon_j^2|x_i, x_j], \end{aligned} \tag{5.4}$$

for $l, l' = 1, ..., p$. Then, we can similarly define critical values as in (3.3) to construct confidence intervals for θ_0 uniform in $K \in \mathcal{K}_n$ analogous to the nonparametric setup. Let

$$\widehat{c}_{1-\alpha} \equiv (1-\alpha) \text{ quantile of } \max_{m=1,...,p} |\widehat{Z}_m|, \quad \widehat{Z}_\Sigma = (\widehat{Z}_1, ..., \widehat{Z}_p)' \sim N(0, \widehat{\Sigma}_n) \tag{5.5}$$

where $\widehat{\Sigma}_n$ is a consistent estimator for unknown Σ defined in (5.4).

Theorem 5.1 is the main result for the partially linear model setup and provides the asymptotic coverage results of the CIs uniform in $K \in \mathcal{K}_n$ analogous to the nonparametric setup in Section 3.

Theorem 5.1. *Suppose that Assumptions 5.1 and 5.2 hold. In addition, assume that* $n\overline{K}^{-2(\gamma_g+\gamma_{g_w})} \to 0$ *and* $\max_{K,K' \in \mathcal{K}_n} |\frac{\widehat{V}_n(K,K')}{V_n(K,K')} - 1| = o_p(1)$ *as* $n, K \to \infty$*. Then,*

$$\lim_{n\to\infty} P(\theta_0 \in [\widehat{\theta}_n(K) \pm \widehat{c}_{1-\alpha}\sqrt{\widehat{V}_n(K)/n}], \quad \forall K \in \mathcal{K}_n) = 1-\alpha, \tag{5.6}$$

$$\liminf_{n\to\infty} P(\theta_0 \in [\widehat{\theta}_n(\widehat{K}) \pm \widehat{c}_{1-\alpha}\sqrt{\widehat{V}_n(\widehat{K})/n}]) \geq 1-\alpha, \quad \widehat{K} \in \mathcal{K}_n, \tag{5.7}$$

where the critical value $\widehat{c}_{1-\alpha}$ *is defined in* (5.5)*.*

Remark 5.1. Note that the construction of CIs requires consistent variance estimation of $\Omega_n(K)$. As discussed in Cattaneo, Jansson, and Newey (2018a, 2018b), the construction of the heteroskedasticity-robust estimator for $\Omega_n(K)$ under $K/n \to c > 0$ is challenging, and the Eicker-Huber-White-type variance estimator generally requires $K/n \to 0$ for consistency. Cattaneo, Jansson, and Newey (2018b) considers the following standard error formula:

$$\widehat{\Omega}_n(K, \kappa_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \kappa_{ij}\hat{v}_{K,i}^2\hat{\varepsilon}_{K,j}^2 \tag{5.8}$$

where $\hat{v}_K = M_K W, \hat{\varepsilon}_K = M_K(Y - W\widehat{\theta}_n(K))$ and symmetric matrix κ_n with (i,j) element κ_{ij}. Cattaneo, Jansson, and Newey (2018b) show that $\widehat{\Omega}_n(K, \kappa_n)$ is consistent even under heteroskedasticity and $K/n \to c > 0$ with a certain choice of κ_n and provide a sufficient condition for consistency. See Theorems 3 and 4 of Cattaneo, Jansson, and Newey (2018b) for further discussion.

6 Simulations

This section investigates the small sample performance of the proposed inference methods. We report the empirical coverage and the average length of the confidence intervals/confidence bands considered in Sections 3 and 4 with various simulation setups.

We consider the following data generating process:

$$y_i = g(x_i) + \varepsilon_i,$$
$$x_i = \Phi(x_i^*), \begin{pmatrix} x_i^* \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2(x_i^*) \end{pmatrix}\right)$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function needed to ensure compact support, and $\sigma^2(x_i^*) = ((1+2x_i^*)/2)^2$ (heteroskedastic). We investigate the following three functions for $g(x)$: $g_1(x) = \ln(|6x-3|+1)sgn(x-1/2)$, $g_2(x) = \frac{\sin(7\pi x/2)}{1+2x^2(sgn(x)+1)}$, and $g_3(x) = x - 1/2 + 5\phi(10(x-1/2))$, where $\phi(\cdot)$ is the standard normal probability density function, and $sgn(\cdot)$ is the sign function. $g_1(x)$ is used in Newey and Powell (2003), as well as Chen and Christensen (2018). $g_2(x)$ and $g_3(x)$ are rescaled versions used in Hall and Horowitz (2013). See Figure 1 for the

shapes of all three functions on $[0,1]$. For all simulation results below, we generate 2000 simulation replications for each design with a sample size $n = 200$.

Results for quadratic splines with evenly placed knots are reported where the number of knots K are selected among $\mathcal{K}_n = \{6, 7, ..., 12\}$ by setting $\underline{K} = 2n^{1/5}$ and $\overline{K} = 2n^{1/3}$ rounded up to the nearest integer. Then, we calculate a pointwise coverage rate (COV) and the average length (AL) of various 95% nominal CIs, as well as analogous uniform CBs for the grid points of x on the support $\mathcal{X} = [0.05, 0.95]$. To calculate critical values, 1000 additional Monte Carlo or bootstrap replications are performed on each simulation iteration. In addition, we investigate results for homoskedastic errors ($\sigma^2(x_i^*) = 1$), different sample sizes $n = \{100, 500\}$, polynomial regressions, and different specifications as in Cattaneo and Farrell (2013) with multivariate and non-normal regressors; however, the results show qualitatively similar patterns and hence are not reported here for brevity. Additional simulation results are reported in the Online Supplementary Material.

Table 1 reports the nominal 95% coverage of the following pointwise CIs at $x = 0.2, 0.5, 0.8, 0.9$: (1) the standard CI in (2.5) with $\widehat{K}_{\texttt{cv}} \in \mathcal{K}_n$ selected to minimize the leave-one-out cross-validation; (2) robust CI in (3.6) with $\widehat{K}_{\texttt{cv}}$ using the critical value $\widehat{c}_{1-\alpha}(x)$; (3) robust CI using $\widehat{K}_{\texttt{cv+}} = \widehat{K}_{\texttt{cv}} + 2$. Analogous uniform inference results for CBs are also reported. The critical values, $\widehat{c}_{1-\alpha}(x)$ and $\widehat{c}_{1-\alpha}$ are constructed using the Monte Carlo methods and weighted bootstrap method, respectively.

Overall, we find that the coverage of the standard CI with $\widehat{K}_{\texttt{cv}}$ is far less than 95% over the support although it has the shortest length. However, the coverage of robust CIs based on $\widehat{K}_{\texttt{cv}}$ or $\widehat{K}_{\texttt{cv+}}$ with $\widehat{c}_{1-\alpha}(x)$ is close to or above 95% and performs well across the different simulation designs, and this is consistent with theoretical results in Theorem 3.1. Using the undersmoothed $\widehat{K}_{\texttt{cv+}}$ (using more terms than the cross-validation) seems to work quite well at most points and for highly nonlinear designs where there exists relatively large bias, e.g., Model 3 ($g_3(x)$) at $x = 0.5$.[7] Uniform coverage rates of confidence bands with selected K seem conservative, and this is due to the large critical values based on weighted bootstrap methods to be uniform in both $K \in \mathcal{K}_n$ and $x \in \mathcal{X}$, including boundary points.

7 Empirical application

In this section, we illustrate inference procedures by revisiting Blomquist and Newey (2002). Understanding how tax policy affects individual labor supply has been a central issue in labor economics (see Hausman (1985) and Blundell and MaCurdy (1999), among many others). Blomquist and Newey (2002) estimate the conditional mean of hours of work given the individual nonlinear budget sets using nonparametric series estimation. They also estimate the wage elasticity of the expected labor supply and find evidence of possible misspecification of the usual parametric model such as maximum likelihood estimation (MLE).

Specifically, Blomquist and Newey (2002) consider the following model by exploiting an additive

[7] The possibly poor coverage property of the standard kernel-based CIs for $g_3(x)$ at the single peak ($x = 0.5$) was also described in Hall and Horowitz (2013, Figure 3).

structure from the utility maximization with piecewise linear budget sets:

$$h_i = g(x_i) + \varepsilon_i, \quad E(\varepsilon_i|x_i) = 0, \tag{7.1}$$

$$g(x_i) = g_1(y_J, w_J) + \sum_{j=1}^{J-1}[g_2(y_j, w_j, \ell_j) - g_2(y_{j+1}, w_{j+1}, \ell_j)], \tag{7.2}$$

where h_i is the hours worked of the ith individual and $x_i = (y_1, \cdots, y_J, w_1, \cdots, w_J, \ell_1, \cdots, \ell_J, J)$ is the budget set, which can be represented by the intercept y_j (non-labor income), slope w_j (marginal wage rates) and the end point ℓ_j of the jth segment in a piecewise linear budget with J segments. Equation (7.2) for the conditional mean function follows from Theorem 2.1 of Blomquist and Newey (2002), and this additive structure substantially reduces the dimensionality issues. To approximate $g(x)$, they consider the power series, $p_k(x) = (y_J^{p_1(k)} w_J^{q_1(k)}, \sum_{j=1}^{J-1} \ell_j^{m(k)}(y_j^{p_2(k)} w_j^{q_2(k)} - y_{j+1}^{p_2(k)} w_{j+1}^{q_2(k)}))$, $p_2(k) + q_2(k) \geq 1$.

From the Swedish "Level of Living" survey in 1973, 1980 and 1990, they pool the data from three waves and use the data for married or cohabiting men of ages 20-60. Changes in the tax system over three different time periods give a large variation in the budget sets. The sample size is $n = 2321$. See Section 5 of Blomquist and Newey (2002) for more detailed descriptions. They estimate the wage elasticity of the expected labor supply

$$E_w = \bar{w}/\bar{h}[\frac{\partial g(w, \cdots, w, \bar{y}, \cdots, \bar{y})}{\partial w}]|_{w=\bar{w}}, \tag{7.3}$$

which is the regression derivative of $g(x)$ evaluated at the mean of the net wage rates $\bar{w}$, virtual income $\bar{y}$ and level of hours $\bar{h}$.

Table 2 is the same table as in Blomquist and Newey (2002, Table 1). They report estimates $\widehat{E}_w$ and standard errors $SE_{\widehat{E}_w}$ with a different number of series terms by adding additional series terms. For example, the estimates in the second row use the term in the first row (1, y_J, w_J) with the additional terms $(\Delta y, \Delta w)$. Here, $\ell^m \Delta y^p w^q$ denotes approximating the term $\sum_j \ell_j^m(y_j^p w_j^q - y_{j+1}^p w_{j+1}^q)$. Blomquist and Newey (2002) also report cross-validation criteria, CV, for each specification. In their formula, series terms are chosen to maximize CV, which minimizes the asymptotic MSE. In addition to their original table, we add the standard 95% CI for each specification, i.e., $CI(K) = \widehat{E}_w(K) \pm 1.96 SE_{\widehat{E}_w}(K)$. In Table 2, it is ambiguous as to which large model (K) can be used for the inference, and we do not have compelling data-dependent methods for selecting one of the large K for the confidence interval to be reported. Here we want to construct CIs that are robust to specification searches.

Figure 2 displays pointwise 95% uniform CIs for $K_m \in \{K_1, K_2, \cdots, K_{11}\}$, where K_m corresponds to each specification in Table 2 with increasing order of series terms, along with the point estimates and standard 95% confidence interval.[8] From Figure 2, we reject a zero wage elasticity of

[8]It is straightforward to construct $\widehat{c}_{1-\alpha}(x)$ using the covariance structure under the homoskedastic error and it only requires estimated variances for different $K \in \mathcal{K}_n$ that are already reported in the table of Blomquist and Newey (2002). Based on 100,000 simulation repetitions, we have $\widehat{c}_{1-\alpha}(x) = 2.503$.

the labor supply for almost all models except $\overline{K}$. Table 2 also reports robust confidence intervals $CI^{\texttt{sup}}_{\widehat{E}_w}(K) = \widehat{E}_w(K) \pm \widehat{c}_{1-\alpha}(x) SE_{\widehat{E}_w}(K)$ with possibly data-dependent $\widehat{K}$ justified by Theorem 3.1 (eq (3.6)). Note that cross-validation chooses $\widehat{K}_{\texttt{cv}} = K_5$, and the standard CI with $\widehat{K}_{\texttt{cv}}$ is [0.0247, 0.0839] and the robust CI is $[0.0165, 0.0921]$. Using $\widehat{K}_{\texttt{cv+}} = K_6$ or $\widehat{K}_{\texttt{cv++}} = K_7$ widens the standard CI, and the robust CIs are $CI^{\texttt{sup}}_{\widehat{E}_w}(\widehat{K}_{\texttt{cv+}}) = [0.0166, 0.1152], CI^{\texttt{sup}}_{\widehat{E}_w}(\widehat{K}_{\texttt{cv++}}) = [0.0070, 0.1186]$.

8 Conclusion

This paper considers nonparametric inference methods given specification searches over different numbers of series terms in the nonparametric series regression model. We provide methods of constructing uniform CIs and confidence bands by adjusting the conventional normal critical value to the critical value based on the supremum of the t-statistics. The critical values can be constructed using simple Monte Carlo simulation or weighted bootstrap methods. Then, we provide an extension of the proposed CIs in the partially linear model setup. Finally, we investigate the finite sample properties of the proposed methods and illustrate uniform CIs in an empirical example of Blomquist and Newey (2002).

While beyond the scope of this paper, there are some potential directions to extend the results established here. First, investigating the coverage property of CIs with data-dependent $\widehat{K}$ using bias-corrected methods is of interest. In particular, it would be of interest to analyze the bias-corrected CI and confidence bands using cross-validation methods combined with the recent results established in Cattaneo, Farrell, and Feng (2019). Second, an extension of the current theory for quantile regression (e.g., Belloni, Chernozhukov, Chetverikov, and Fernández-Val (2019)) or the nonparametric IV setup would be desirable. In the NPIV setup, for example, one can consider pointwise CIs (or uniform confidence bands) that are uniform in pairs of $(K_n, J_n) \in \mathcal{K}_n \times \mathcal{J}_n$ with an additional dimension of the instrument sieve and the number of instruments $J = J_n$. This is a difficult problem, and it would require a distinct theory to address the ill-posed inverse problem as well as two-dimensional choices. We leave these topics for future research.

References

Andrews, D. W. K. (1991a): "Asymptotic Normality of Series Estimators for Nonparametric and Semiparametric Regression Models," *Econometrica*, 59, 307-345.

Andrews, D. W. K. (1991b): "Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors," *Journal of Econometrics*, 47, 359-377.

Armstrong, T. B. and M. Kolesár (2018): "A Simple Adjustment for Bandwidth Snooping," *Review of Economic Studies*, 85, 732-765.

Belloni, A., V. Chernozhukov, D. Chetverikov, and I. Fernández-Val (2019): "Conditional quantile processes based on series or many regressors," *Journal of Econometrics*, 213, 4-29.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345-366.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608-650.

Blomquist, S. and W. K. Newey (2002): "Nonparametric Estimation with Nonlinear Budget Sets," *Econometrica*, 70, 2455-2480.

Blundell, R. and T. E. MaCurdy (1999): "Labor Supply: A Review of Alternative Approaches," *Handbook of Labor Economics*, In: O. Ashenfelter, D. Card (Eds.), vol. 3., Elsevier, Chapter 27.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113, 767-779.

Cattaneo, M. D. and M. H. Farrell (2013): "Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators," *Journal of Econometrics*, 174, 127-143.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2019): "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, forthcoming.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018a): "Alternative Asymptotics and the Partially Linear Model with Many Regressors," *Econometric Theory*, 34, 277-301.

Cattaneo, M. D., M. Jansson, and W. K. Newey (2018b): "Inference in Linear Regression Models with Many Covariates and Heteroscedasticity," *Journal of the American Statistical Association*, 113, 1350-1361.

Chao, J. C., N. R. Swanson, J. A. Hausman, W. K. Newey, and T. Woutersen (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 28, 42-86.

Chatterjee, S. (2005): "An error bound in the Sudakov-Fernique inequality," arXiv:math/0510424

Chen, X. (2007): "Large Sample Sieve Estimation of Semi-nonparametric Models," *Handbook of Econometrics*, In: J.J. Heckman, E. Leamer (Eds.), vol. 6B., Elsevier, Chapter 76.

Chen, X. and T. Christensen (2015): "Optimal Uniform Convergence Rates and Asymptotic Normality for Series Estimators Under Weak Dependence and Weak Conditions," *Journal of Econometrics*, 188, 447-465.

Chen, X. and T. Christensen (2018): "Optimal Sup-norm Rates and Uniform Inference on Nonlinear Functionals of Nonparametric IV Regression", *Quantitative Economics*, 9(1), 39-85.

Chen, X. and Z. Liao (2014): "Sieve M inference on irregular parameters," *Journal of Econometrics*,182, 70-86.

Chen, X., Z. Liao, and Y. Sun (2014): "Sieve inference on possibly misspecified semi-nonparametric time series models," *Journal of Econometrics*,178, 639-658.

Chen, X. and X. Shen (1998): "Sieve extremum estimates for weakly dependent data," *Econometrica*, 66 (2), 289-314.

Chernozhukov V, D. Chetverikov, and K. Kato (2014a): "Gaussian approximation of suprema of empirical processes," *The Annals of Statistics*, 42(4), 1564-1597.

Chernozhukov V, D. Chetverikov, and K. Kato (2014b): "Anti-Concentration and Honest, Adaptive Confidence Bands," *The Annals of Statistics*, 42(5), 1787-1818.

Chernozhukov V, D. Chetverikov, and K. Kato (2016): "Empirical and multiplier bootstraps for suprema of empirical processes of increasing complexity, and related Gaussian couplings," *Stochastic Processes and their Applications*, 126(12), 3632-3651.

Donald, S. G. and W. K. Newey (1994): "Series Estimation of Semilinear Models," *Journal of Multivariate Analysis*, 50, 30-40.

Eastwood, B. J. and A.R. Gallant, (1991): "Adaptive Rules for Seminonparametric Estimators That Achieve Asymptotic Normality," *Econometric Theory*, 7, 307-340.

Giné, E. and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38, 1122-1170.

Giné, E. and R. Nickl (2015): *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Cambridge University Press.

Hall, P. and J. Horowitz (2013): "A Simple Bootstrap Method for Constructing Nonparametric Confidence Bands for Functions," *The Annals of Statistics*, 41, 1892-1921.

Hansen B. E. (2015): "The Integrated Mean Squared Error of Series Regression and a Rosenthal Hilbert-Space Inequality," *Econometric Theory*, 31, 337-361.

Hansen, P.R. (2005): "A Test for Superior Predictive Ability," *Journal of Business and Economic Statistics*, 23, 365-380.

Härdle, W. and O. Linton (1994): "Applied Nonparametric Methods," *Handbook of Econometrics*, In: R. F. Engle, D. F. McFadden (Eds.), vol. 4., Elsevier, Chapter 38.

Hausman, J. A. (1985): "The Econometrics of Nonlinear Budget Sets", *Econometrica*, 53, 1255-1282.

Heckman, J. J., L. J. Lochner, and P. E. Todd (2006): "Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," *Handbook of the Economics of Education*, In: E. A. Hanushek, and F. Welch (Eds.), Vol. 1, Elsevier, Chapter 7.

Horowitz, J. L. (2014): "Adaptive Nonparametric Instrumental Variables Estimation: Empirical Choice of the Regularization Parameter," *Journal of Econometrics*, 180, 158-173.

Horowitz, J. L. and S. Lee (2012): "Uniform Confidence Bands for Functions Estimated Nonparametrically with Instrumental Variables," *Journal of Econometrics*, 168, 175-188.

Huang, J. Z. (2003): "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600-1635.

Kozbur, D. (2018): "Inference in Additively Separable Models With a High-Dimensional Set of Conditioning Variables," Working Paper, arXiv:1503.05436.

Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics," *The American Economic Review*, 73, 31-43.

Lepski, O. V. (1990): "On a problem of adaptive estimation in Gaussian white noise," *Theory of Probability and its Applications*, 35, 454-466.

Li, K. C. (1987): "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15, 958-975.

Li, Qi, and J. S. Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Linton, O. (1995): "Second order approximation in the partialy linear regression model," *Econometrica*, 63(5), 1079-1112.

Newey, W. K. (1994a): "Series Estimation of Regression Functionals," *Econometric Theory*, 10, 1-28.

Newey, W. K. (1994b): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349-1382.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147-168.

Newey, W. K. (2013): "Nonparametric Instrumental Variables Estimation," *American Economic Review: Papers & Proceedings*, 103, 550-556.

Newey, W. K. and J. L. Powell (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565-1578.

Newey, W. K. and J. L. Powell, F. Vella (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565-603.

Robinson, P. M. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56(4), 931-954.

Romano, J. P. and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73, 1237-1282.

Schennach, S. M. (2015): "A bias bound approach to nonparametric inference," *CEMMAP working paper CWP71/15.*

Van Der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68, 1097-1126.

Zhou, S., X. Shen, and D.A. Wolfe (1998): "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760-1782.

Appendix A Proofs

A.1 Preliminaries and Useful Lemmas

We define additional notations for the empirical process theory used in the proof of Theorem 4.1. Given measurable space $(S, \mathcal{S})$, let $\mathcal{F}$ as a class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. For any probability measure Q on $(S, \mathcal{S})$, we define $N(\epsilon, \mathcal{F}, L_2(Q))$ as covering numbers, which is the minimal number of the $L_2(Q)$ balls of radius ϵ to cover $\mathcal{F}$ with $L_2(Q)$ norms $||f||_{Q,2} = (\int |f|^2 dQ)^{1/2}$. The uniform entropy numbers relative to the $L_2(Q)$ norms are defined as $\sup_Q \log N(\epsilon||F||_{Q,2}, \mathcal{F}, L_2(Q))$ where the supremum is over all discrete probability measures with an envelope function F. We define $\mathcal{F}$ as a *VC type* with envelope F if there are constants $A, v > 0$ such that $\sup_Q N(\epsilon||F||_{Q,2}, \mathcal{F}, L_2(Q)) \leq (A/\epsilon)^v$ for all $0 < \epsilon \leq 1$.

Let the data $z_i = (\varepsilon_i, x_i)$ be i.i.d. random vectors defined on the probability space ($\mathcal{Z} = \mathcal{E} \times \mathcal{X}$, $\mathcal{A}, P$) with common probability distribution $P \equiv P_{\varepsilon,x}$. We think of $(\varepsilon_1, x_1), \cdots (\varepsilon_n, x_n)$ as the coordinates of the infinite product probability space. We avoid discussing nonmeasurability issues and outer expectations (for the related issues, see van der Vaart and Wellner (1996)). Throughout the proofs, we denote $c, C > 0$ as universal constants that do not depend on n.

For any sequence $\{K = K_n : n \geq 1\} \in \prod_{n=1}^{\infty} \mathcal{K}_n$ under Assumption 2.1, we first define the orthonormalized vector of basis functions

$$\tilde{P}(K,x) \equiv Q_K^{-1/2} P(K,x) = E[P_{Ki}P_{Ki}']^{-1/2} P(K,x),\ \tilde{P}_{Ki} = \tilde{P}(K,x_i),\ \tilde{P}^K = [\tilde{P}_{K1}, \cdots, \tilde{P}_{Kn}]'.$$

We observe that

$$\widehat{g}_n(K,x) = \tilde{P}(K,x)'(\tilde{P}^{K\prime}\tilde{P}^K)^{-1}\tilde{P}^{K\prime}Y,\quad V_n(K,x) = \tilde{P}(K,x)'\tilde{\Omega}_K\tilde{P}(K,x),\quad \tilde{\Omega}_K = E(\tilde{P}_{Ki}\tilde{P}_{Ki}'\varepsilon_i^2).$$

Without loss of generality, we may impose normalizations of $Q_{\overline{K}} = I_{\overline{K}}$ or $Q_K = E(P_{Ki}P_{Ki}') = I_K$ uniformly over $K \in \mathcal{K}_n$, since $\widehat{g}_n(K,x)$ is invariant to nonsingular linear transformations of $P(K,x)$. However, we shall treat Q_K as unknown and deal with the non-orthonormalized series terms. Next, we re-define pseudo true value β_K, with an abuse of notation, using orthonormalized series terms $\tilde{P}_{Ki}$. That is, $y_i = \tilde{P}_{Ki}'\beta_K + \varepsilon_{Ki}, E[\tilde{P}_{Ki}\varepsilon_{Ki}] = 0$ where $\varepsilon_{Ki} = r_{Ki} + \varepsilon_i$, $r_n(K,x) = g_0(x) - \tilde{P}(K,x)'\beta_K$, $r_{Ki} = r_n(K,x_i)$, and $r_K \equiv (r_{K1}, \cdots r_{Kn})'$. We also define $\widehat{Q}_K \equiv \frac{1}{n}\tilde{P}^{K\prime}\tilde{P}^K$, $\underline{\sigma}^2 \equiv \inf_x E[\varepsilon_i^2|x_i = x]$, $\bar{\sigma}^2 \equiv \sup_x E[\varepsilon_i^2|x_i = x]$.

We first provide useful lemmas which will be used in the proof of Theorem 3.1 and 4.1. The versions of proof of Lemmas 1 and 2 with $\mathcal{K}_n = \{K\}$ are available in the literature, such as Belloni et al. (2015) and Chen and Christensen (2015), among many others. The maximal inequalities are used in the proof of Lemmas 1 and 2 to bound the remainder terms in the linearization of the t-statistics. Also note that different rate conditions of K such as those in Newey (1997) can be used here but lead to different bounds. We provide the proofs of Lemma 1 and 2 in the Online Supplementary Material (Section B).

Lemma 1. *Suppose that Assumptions 2.1, 3.1, and 3.2 hold, then* $||\widehat{Q}_K - I_K|| = O_p(\sqrt{\lambda_K^2 \zeta_K^2 \log K/n})$ *for any* $K \in \mathcal{K}_n$ *and the following holds*

$$\max_{K\in\mathcal{K}_n} |R_1(K,x)| = O_p(\max_{K\in\mathcal{K}_n} \sqrt{\frac{\lambda_K^2\zeta_K^2 \log K \log p}{n}}(1+\ell_K c_K\sqrt{K})), \tag{A.1}$$

$$\max_{K\in\mathcal{K}_n} |R_2(K,x)| = O_p(\max_{K\in\mathcal{K}_n} (\ell_K c_K)\sqrt{\log p}), \tag{A.2}$$

where $R_1(K,x) \equiv \sqrt{\frac{1}{nV_n(K,x)}}\tilde{P}(K,x)'(\widehat{Q}_K^{-1} - I_K)\tilde{P}^{K\prime}(\varepsilon + r_K)$, $R_2(K,x) \equiv \sqrt{\frac{1}{nV_n(K,x)}}\tilde{P}(K,x)'\tilde{P}^{K\prime} r_K$.

Lemma 2. *Suppose that Assumptions 2.1, 3.1 and 4.1 hold, then the following holds*

$$\sup_{K\in\mathcal{K}_n, x\in\mathcal{X}} |R_1(K,x)| = O_p(\max_{K\in\mathcal{K}_n} \sqrt{\frac{\lambda_K^2\zeta_K^2 \log K \log n}{n}}(n^{1/q} + \ell_K c_K\sqrt{K})), \tag{A.3}$$

$$\sup_{K\in\mathcal{K}_n, x\in\mathcal{X}} |R_2(K,x)| = O_p(\max_{K\in\mathcal{K}_n} (\ell_K c_K)\sqrt{\log n}), \tag{A.4}$$

where $R_1(K,x), R_2(K,x)$ *are defined in Lemma 1.*

A.2 Proofs of the Main Results

A.2.1 Proof of Theorem 3.1

Proof. For any $K \in \mathcal{K}_n$, we first consider the decomposition of the t-statistic in (2.4) with the known variance $V_n(K,x)$,

$$\begin{aligned} T_n(K,x) &= \sqrt{\frac{n}{V_n(K,x)}}\tilde{P}(K,x)'(\widehat{\beta}_K - \beta_K) - \sqrt{\frac{n}{V_n(K,x)}} r_n(K,x) \\ &= t_n(K,x) + R_1(K,x) + R_2(K,x) + \nu_n(K,x) \end{aligned}$$

where $t_n(K,x) = n^{-1/2}\sum_{i=1}^n \frac{\tilde{P}(K,x)'\tilde{P}_{Ki}\varepsilon_i}{V_n(K,x)^{1/2}}$, $R_1(K,x), R_2(K,x)$ are defined in Lemma 1, and $\nu_n(K,x) = -\sqrt{n}V_n(K,x)^{-1/2} r_n(K,x)$. Define

$$t_n \equiv (t_n(K_1,x), \cdots, t_n(K_p,x))' = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$$

where $\xi_i = (\xi_{i1}, \xi_{i2}, \cdots, \xi_{ip})' \in \mathbb{R}^p$ with $\xi_{ij} = \frac{\tilde{P}(K_j,x)'\tilde{P}_{K_j i}\varepsilon_i}{V_n(K_j,x)^{1/2}}$ and $p = |\mathcal{K}_n|$. Note that $E[\xi_{ij}] = 0$ and $E[|\xi_{ij}|^3] \lesssim E[|\tilde{P}(K_j,x)'\tilde{P}_{K_j i}/V_n(K_j,x)^{1/2}|^3] \sup_x E[|\varepsilon_i|^3 | x_i = x] \lesssim \max_K \zeta_K$ for all $1 \le i \le n$, $1 \le j \le p$. By Lemma A.2 in the Online Supplementary Material, for any $\delta > 0$, there exists a random variable $\max_{1\le j\le p}\sum_{i=1}^n Z_{ij}$ with independent random vectors $\{Z_i\}_{i=1}^n \in \mathbb{R}^p$, $Z_i \sim N(0, \frac{1}{n}E[\xi_i\xi_i'])$, $1 \le i \le n$, such that

$$P(|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|| > 16\delta) \lesssim \frac{\log(p\vee n)}{\delta^2} D_1 + \frac{\log^2(p\vee n)}{\delta^3 n^{3/2}}(D_2 + D_3) + \frac{\log n}{n}$$

where $D_1 = E\big[\max_{1\le j,l\le p}|\frac{1}{n}\sum_{i=1}^n(\xi_{ij}\xi_{il} - E[\xi_{ij}\xi_{il}])|\big]$, $D_2 = E\big[\max_{1\le j\le p}\sum_{i=1}^n|\xi_{ij}|^3\big]$, and $D_3 = \sum_{i=1}^n E\big[\max_{1\le j\le p}|\xi_{ij}|^3 1\big(\max_{1\le j\le p}|\xi_{ij}| > \delta\sqrt{n}/\log(p\vee n)\big)\big]$.

First consider the case (a) in Assumption 3.2(ii). Combining bounds for D_1, D_2, D_3 in Lemma B.1 in the Online Supplementary Material gives, for any $\delta > 0$,

$$
\begin{aligned}
&P(|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|| > 16\delta)\\
&\lesssim \frac{\log(p\vee n)}{\delta^2}\big[(\frac{(\max_K\zeta_K)^2\log p}{n})^{1/2} + \frac{(\max_K\zeta_K)^2\log p}{n^{1-2/q}}\big]\\
&+\frac{\log^2(p\vee n)}{\delta^3}\big[(\frac{(\max_K\zeta_K)^2}{n})^{1/2} + \frac{(\max_K\zeta_K)^3\log p}{n^{3/2-3/q}}\big] + \frac{\log^{q-1}(p\vee n)}{\delta^q}\frac{(\max_K\zeta_K)^q}{n^{q/2-1}} + \frac{\log n}{n}.
\end{aligned}
$$

For $\gamma > 0$, by setting

$$
\begin{aligned}
\delta = \quad & \gamma^{-1/3}\big(\frac{(\max_K\zeta_K)^2\log^4(p\vee n)}{n}\big)^{1/6} + \gamma^{-1/2}\big(\frac{(\max_K\zeta_K)^2\log(p\vee n)\log p}{n^{1-2/q}}\big)^{1/2}\\
&+\gamma^{-1/3}\big(\frac{(\max_K\zeta_K)^3\log^2(p\vee n)\log p}{n^{3/2-3/q}}\big)^{1/3},
\end{aligned}
$$

we have

$$
P(|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|| > C_1\delta) \le C_2(\gamma + \frac{\log n}{n})
$$

where C_1, C_2 are positive constants that depend only on q. If we take $\gamma = \gamma_n \to 0$ sufficiently slowly, e.g., $\gamma = \log(p\vee n)^{-1/2}$, then the above implies there exists $\max_{1\le j\le p}\sum_{i=1}^n Z_{ij}$ such that

$$
|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|| = o_p((\frac{(\max_K\zeta_K)^2\log^5(p\vee n)}{n})^{1/6} + \frac{(\max_K\zeta_K)\log^{3/4}(p\vee n)\log^{1/2}p}{n^{1/2-1/q}}).
$$

Next, consider the case (b) in Assumption 3.2(ii). For any $\delta > 0$,

$$
\begin{aligned}
&P(|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|| > 16\delta)\\
&\lesssim \frac{\log(p\vee n)}{\delta^2}\big[(\frac{(\max_K\zeta_K)^2\log p}{n})^{1/2} + \frac{(\max_K\zeta_K)^2\log^2(pn)\log p}{n}\big]\\
&+\frac{\log^2(p\vee n)}{\delta^3}\big[(\frac{(\max_K\zeta_K)^2}{n})^{1/2} + \frac{(\max_K\zeta_K)^3\log^3(pn)\log p}{n^{3/2}}\big]\\
&+\frac{\log^2(p\vee n)}{\delta^3}\big[\frac{1}{n^{1/2}}(\frac{\delta^3 n^{3/2}}{\log^3(p\vee n)} + (\max_K\zeta_K)^3\log^3 p)\exp(-\frac{\delta\sqrt{n}}{C\max_K\zeta_K\log p\log(p\vee n)})\big] + \frac{\log n}{n}
\end{aligned}
$$

by Lemma B.1 in the Online Supplementary Material. Similarly, by setting

$$
\delta = \max\{\gamma^{-1/3}(\max_K\zeta_K)^2\log^4(p\vee n)/n)^{1/6}, 2C((\max_K\zeta_K)^2\log^4(p\vee n)\log^2 p/n)^{1/2}\}
$$

we have, for $\gamma > 0$,

$$P(|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|| > C_1\delta) \le C_2(\gamma + \frac{\log n}{n})$$

where C_1, C_2 are universal constants which do not depend on n. Here we use $\frac{\delta\sqrt{n}}{C\max_K \zeta_K \log p \log(p\vee n)} \ge 2\log(p\vee n)$. By taking $\gamma = \log(p\vee n)^{-1/2}$, there exists $\max_{1\le j\le p}\sum_{i=1}^n Z_{ij}$ such that

$$|\max_{1\le j\le p}|t_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|| = o_p((\frac{(\max_K \zeta_K)^2\log^5(p\vee n)}{n})^{1/6} + \frac{(\max_K \zeta_K)^2\log^4(p\vee n)\log^2 p}{n})^{1/2}).$$

In either case (a) or (b), the above coupling inequality shows that there exists a sequence of random variables $\max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|$ such that $\big|\max_{K\in\mathcal{K}_n}|t_n(K,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\big| = o_p(a_n)$, $a_n = 1/(\log p)^{1/2}$ under the rate conditions imposed in Theorem 3.1. Furthermore,

$$\begin{aligned}\big|\max_{1\le j\le p}|T_n(K_j,x)| - \max_{1\le j\le p}|t_n(K_j,x)|\big| &\le \max_{1\le j\le p}|T_n(K_j,x) - t_n(K_j,x)| \le \max_{1\le j\le p}|R_1(K_j,x)| \\ &\quad + \max_{1\le j\le p}|R_2(K_j,x)| + \max_{1\le j\le p}|\nu_n(K_j,x)| = o_p(a_n) \quad \text{(A.5)}\end{aligned}$$

with $a_n = 1/(\log p)^{1/2}$ by Lemma 1 and the assumption imposed in Theorem 3.1. We also have

$$\begin{aligned}&\big|\max_{1\le j\le p}|T_n(K_j,x)| - \max_{1\le j\le p}|\widehat{T}_n(K_j,x)|\big| \le \max_{1\le j\le p}|T_n(K_j,x) - \widehat{T}_n(K_j,x)| \\ &\le \max_{1\le j\le p}|T_n(K_j,x)|\max_{1\le j\le p}|1 - \frac{V_n(K_j,x)^{1/2}}{\widehat{V}_n(K_j,x)^{1/2}}| = o_p(a_n) \quad \text{(A.6)}\end{aligned}$$

where we use Lemma 1 and $\max_{1\le j\le p}|t_n(K_j,x)| \lesssim_P \sqrt{\log p}$ by the maximal inequality (e.g., Lemma A.4 in the Online Supplementary Material) and Assumption 3.2(iii) with $a_n = 1/(\log p)^{1/2}$. Combining (A.5) and (A.6) gives $\big|\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\big| = o_p(a_n)$ with $a_n = 1/(\log p)^{1/2}$. Then, there exists some sequence of positive constant δ_n such that $\delta_n = o(1)$ and $P(\big|\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\big| > a_n\delta_n) = o(1)$.

For any $u \in \mathbb{R}$, we have

$$\begin{aligned}&P(\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| \le u) \\ &\le P(\{\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| \le u\} \cap \{\big|\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\big| \le a_n\delta_n\}) \\ &\quad + P(\big|\max_{1\le j\le p}|\widehat{T}_n(K_j,x)| - \max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|\big| > a_n\delta_n) \\ &\le P(\max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}| \le u + a_n\delta_n) + o(1) \le P(\max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}| \le u) + a_n\delta_n E[\max_{1\le j\le p}\sum_{i=1}^n |Z_{ij}|] + o(1)\end{aligned}$$

where the last inequality uses anti-concentration inequality (Lemma A.8 in the Online Supplemen-

tary Material). The reverse inequality holds with a similar argument above, and thus

$$\sup_{u\in\mathbb{R}}\big|P(\max_{1\le j\le p}|\widehat{T}_n(K,x)|\le u)-P(\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|\le u)\big|=a_n\delta_nE[\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|]+o(1)=o(1)$$

where we use $E[\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|]\lesssim\sqrt{\log p}$ by Gaussian maximal inequality and $a_n=(\log p)^{-1/2}$. Using the same arguments above, $\big|\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|-\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}|\big|=o_p(a_n)$ by Sudakov-Fernique type bound (e.g., Chatterjee (2005)) and Assumption 3.2(iii), we have $\sup_{u\in\mathbb{R}}\big|P(\max_{1\le j\le p}|\widehat{Z}_{ij}||\le u)-P(\max_{1\le j\le p}\sum_{i=1}^n|Z_{ij}|\le u)|=o(1)$. Therefore, the following holds by the triangle inequality,

$$\sup_{u\in\mathbb{R}}\big|P(\max_{1\le j\le p}|\widehat{T}_n(K,x)|\le u)-P(\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}|\le u)\big|=o(1),$$

and then we conclude

$$P(\max_{K\in\mathcal{K}_n}|\widehat{T}_n(K,x)|\le\widehat{c}_{1-\alpha}(x))=1-\alpha+o(1),$$

with a critical value $\widehat{c}_{1-\alpha}(x)$ given in (3.3), and the coverage result (3.5) follows.

Finally, we will show (3.6). For $\widehat{K}\in\mathcal{K}_n$, observe that

$$|\widehat{T}_n(\widehat{K},x)|\le(|t_n(\widehat{K},x)|+|R_1(\widehat{K},x)|+|R_2(\widehat{K},x)|+|\nu_n(\widehat{K},x)|)|\frac{V_n(\widehat{K},x)^{1/2}}{\widehat{V}_n(\widehat{K},x)^{1/2}}| \tag{A.7}$$

by the triangle inequality. Then,

$$\begin{aligned}
&P(g_0(x)\in[\widehat{g}_n(\widehat{K},x)\pm\widehat{c}_{1-\alpha}(x)\sqrt{\widehat{V}_n(\widehat{K},x)/n}])\\
&\ge P(|t_n(\widehat{K},x)|+|R_1(\widehat{K},x)|+|R_2(\widehat{K},x)|+|\nu_n(\widehat{K},x)|\le\widehat{c}_{1-\alpha}(x)|\frac{\widehat{V}_n(\widehat{K},x)^{1/2}}{V_n(\widehat{K},x)^{1/2}}|)\\
&\ge P(|t_n(\widehat{K},x)|+|R_1(\widehat{K},x)|+|R_2(\widehat{K},x)|+|\nu_n(\widehat{K},x)|\le\widehat{c}_{1-\alpha}(x)(1-a_n^2\delta_{1n}))-\epsilon_{1n} &\text{(A.8)}\\
&\ge P(|t_n(\widehat{K},x)|\le\widehat{c}_{1-\alpha}(x)(1-a_n^2\delta_{1n})-a_n\delta_{2n}-a_n\delta_{3n})-\epsilon_{1n}-\epsilon_{2n}-\epsilon_{3n} &\text{(A.9)}\\
&\ge P(\max_{K\in\mathcal{K}_n}|t_n(K,x)|\le\widehat{c}_{1-\alpha}(x)(1-a_n^2\delta_{1n})-a_n\delta_{2n}-a_n\delta_{3n})-\epsilon_{1n}-\epsilon_{2n}-\epsilon_{3n} &\text{(A.10)}\\
&\ge P(\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}|\le\widehat{c}_{1-\alpha}(x)-\tilde{\delta}_n)-\tilde{\epsilon}_n &\text{(A.11)}\\
&\ge 1-\alpha-\sup_u P(|\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}|-u|\le\tilde{\delta}_n)-\tilde{\epsilon}_n\ \ge 1-\alpha-o(1). &\text{(A.12)}
\end{aligned}$$

The first inequality follows by (A.7), and (A.8) holds by Assumption 3.2(iii) with some sequence of positive constant $\delta_{1n}=o(1),\epsilon_{1n}=o(1)$ and (A.9) follows by $|R_1(\widehat{K},x)|+|R_2(\widehat{K},x)|=o_p(a_n)$ from Lemma 1 and the assumption $|\frac{\sqrt{n}r_n(\widehat{K},x)}{V_n(\widehat{K},x)^{1/2}}|=o(a_n)$ with $a_n=1/(\log p)^{1/2}$ and some sequences of constants $\delta_{2n}=o(1),\epsilon_{2n}=o(1),\delta_{3n}=o(1),\epsilon_{3n}=o(1)$. (A.10) follows by $|t_n(\widehat{K},x)|\le\max_{K\in\mathcal{K}_n}|t_n(K,x)|$, and (A.11) holds by $|\max_{K\in\mathcal{K}_n}|t_n(K,x)|-\max_{1\le j\le p}\sum_{i=1}^n|\widehat{Z}_{ij}||=o_p(a_n)$ with

some sequences $\delta_{4n} = o(1), \epsilon_{4n} = o(1)$ and defining $\tilde{\delta}_n = \widehat{c}_{1-\alpha}(x)a_n^2\delta_{1n} + a_n\delta_{2n} + a_n\delta_{3n} + a_n\delta_{4n}$, $\tilde{\epsilon}_n = \epsilon_{1n} + \epsilon_{2n} + \epsilon_{3n} + \epsilon_{4n}$. Finally, (A.12) holds by Lemma A.8, $E[\max_{1\leq j\leq p}\sum_{i=1}^n |\widehat{Z}_{ij}|] \lesssim \sqrt{\log p}$ and $\tilde{\delta}_n\sqrt{\log p} = o(1)$ since $\widehat{c}_{1-\alpha}(x) \lesssim \sqrt{\log p}$ by Lemma A.15. This completes the proof. ■

A.2.2 Proof of Theorem 4.1

Proof. Similar to the proof of Theorem 3.1, we have the following linearization of the t-statistics uniformly in $(K, x) \in \mathcal{K}_n \times \mathcal{X}$,

$$T_n(K, x) = t_n(K, x) + \nu_n(K, x) + R_n(K, x),$$

where $t_n(K, x) = n^{-1/2}\sum_{i=1}^n \tilde{P}(K, x)'\tilde{P}_{Ki}\varepsilon_i/V_n(K, x)^{1/2}$ and $R_n(K, x) = R_1(K, x) + R_2(K, x)$. Define $f_{n,K,x} : (\mathcal{E} \times \mathcal{X}) \mapsto \mathbb{R}$ for given $n \geq 1$, $K \in \mathcal{K}_n, x \in \mathcal{X}$,

$$f_{n,K,x}(\varepsilon, t) = \frac{\tilde{P}(K, x)'\tilde{P}(K, t)\varepsilon}{V_n(K, x)^{1/2}}, (\varepsilon, t) \in \mathcal{E} \times \mathcal{X}. \tag{A.13}$$

and consider the class of measurable functions $\mathcal{F}_n = \{f_{n,K,x} : (K, x) \in \mathcal{K}_n \times \mathcal{X}\}$. Then, we consider the following empirical process:

$$\Big\{t_n(K, x) : (K, x) \in \mathcal{K}_n \times \mathcal{X}\Big\} = \Big\{n^{-1/2}\sum_{i=1}^n f_{n,K,x}(\varepsilon_i, x_i) : (K, x) \in \mathcal{K}_n \times \mathcal{X}\Big\}$$

which is indexed by classes of functions $\mathcal{F}_n$. Define $\alpha(K, x) \equiv \tilde{P}(K, x)/V_n(K, x)^{1/2} = \tilde{P}(K, x)/||\Omega_K^{1/2}\tilde{P}(K, x)||$. Note that $|f_{n,K,x}(\varepsilon, t)| = |\alpha(K, x)'\tilde{P}(K, t)\varepsilon| \leq C|\varepsilon| \max_K \zeta_K$ for any $(K, x) \in \mathcal{K}_n \times \mathcal{X}$. We define the envelope function $F_n(\varepsilon, t) \equiv C|\varepsilon| \max_K \zeta_K \vee 1$. By Assumption 4.1, we have

$$\begin{aligned}
&|f_{n,K,x} - f_{n,K',x'}| = |\varepsilon||\alpha(K, x)'\tilde{P}(K, t) - \alpha(K', x')'\tilde{P}(K', t)| \\
&\leq |\varepsilon|\big|\big[|\alpha(K, x)'\tilde{P}(K, t) - \alpha(K, x)'\tilde{P}(K', t)| + |\alpha(K, x)'\tilde{P}(K', t) - \alpha(K', x)'\tilde{P}(K', t)| \\
&\quad + |\alpha(K', x)'\tilde{P}(K', t) - \alpha(K', x')'\tilde{P}(K', t)|\big] \leq |\varepsilon| A \max_K \zeta_K L_n(||x - x'|| + |K - K'|)
\end{aligned}$$

for all $x, x' \in \mathcal{X}, K, K' \in \mathcal{K}_n$ where $L_n = \zeta^{L_1} \vee \zeta^{L_2}$. Therefore, the class of functions $\mathcal{F}_n = \{f_{n,K,x} : (K, x) \in \mathcal{K}_n \times \mathcal{X}\}$ is a VC type and there are constants $A, V > 0$ such that

$$\sup_Q N(\epsilon||F_n||_{L^2(Q)}, \mathcal{F}_n, L^2(Q)) \leq (AL_n/\epsilon)^V, 0 < \forall\epsilon \leq 1$$

for each n. Then, using Theorem 2.1 (Lemma A.9 in the Online Supplementary Material) in Chernozhukov et al. (2016) with $B(f) = 0$, there exists a tight Gaussian process $G_n(f)$ in $\ell^\infty(\mathcal{F}_n)$ and $Z_n(K, x) = G_n(f_{n,K,x})$ in $\ell^\infty(\mathcal{K}_n \times \mathcal{X})$ with zero mean and covariance function (4.2), $E[G_n(f)G_n(f')] = Cov(f_{n,K,x}(\varepsilon_i, x_i), f'_{n,K',x'}(\varepsilon_i, x_i))$ and a sequence of random variables

$\widetilde{Z} \equiv \sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |Z_n(K,x)|$ such that, for every $\gamma \in (0,1)$,

$$P(|\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |t_n(K,x)| - \widetilde{Z}| > C_1\delta_{1n}) \leq C_2(\gamma + n^{-1}) \tag{A.14}$$

where C_1, C_2 are positive constants that depend only on q, and

$$\delta_{1n} = \gamma^{-1/q} n^{-1/2+1/q} \max_K \zeta_K \log n + \gamma^{-1/3} n^{-1/6} (\max_K \zeta_K)^{1/3} \log^{2/3} n$$

by Assumption 4.1(iii) and assuming $\log^3 n \leq n$. By taking $\gamma = (\log n)^{-1/2}$, we have

$$|\sup_{K,x} |t_n(K,x)| - \widetilde{Z}| = o_p(n^{-1/2+1/q} \max_K \zeta_K \log^{1+1/2q} n + n^{-1/6} (\max_K \zeta_K)^{1/3} \log^{5/6} n).$$

Furthermore, $|R_1(K,x)| = o_p(a_n), |R_2(K,x)| = o_p(a_n), |\nu_n(K,x)| = o_p(a_n)$ uniformly in $(K, x) \in \mathcal{K}_n \times \mathcal{X}$ with $a_n = 1/(\log n)^{1/2}$ by Lemma 2 and the rate conditions. Again, consider the class of functions $\mathcal{F}_n = \{f_{n,K,x} : (K,x) \in \mathcal{K}_n \times \mathcal{X}\}$ and then

$$E\big[\sup_{K,x} |t_n(K,x)|\big] \lesssim \sqrt{\log n} + (\max_K \zeta_K)^{q/(q-2)} \log n/\sqrt{n} \lesssim \sqrt{\log n}$$

by Lemma A.13 and Assumption 4.1(iii), and we have $\sup_{K,x} |t_n(K,x)| \lesssim_P \sqrt{\log n}$. Further, $\sup_{K,x} |Z_n(K,x)| \lesssim_P \sqrt{\log n}$ using Dudley's inequality (Corollary 2.2.8 in van der Vaart and Wellner (1996)) and using the same arguments given in Theorem 3.1, we have $\big|\sup_{K,x} |\widehat{T}_n(K,x)| - \widetilde{Z}\big| = o_p(a_n)$ with $a_n = 1/(\log n)^{1/2}$ and

$$\sup_{u\in\mathbb{R}} \big|P(\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n(K,x)| \leq u) - P(\widetilde{Z} \leq u)\big| = o(1). \tag{A.15}$$

Next we consider following (infeasible) bootstrap process

$$T_n^e(K,x) = \frac{\sqrt{n}(\widehat{g}_n^e(K,x) - \widehat{g}_n(K,x))}{V_n(K,x)^{1/2}}, \quad (K,x) \in \mathcal{K}_n \times \mathcal{X}$$

where $\widehat{g}_n^e(K,x) = \tilde{P}(K,x)'\widehat{\beta}_K^e$, $\widehat{\beta}_K^e$ is defined in (4.4) with $\tilde{P}(K,x_i)$, and e_i is i.i.d. standard exponential random variables independent of $X^n = \{x_1, ..., x_n\}$. Then, we have

$$\begin{aligned} T_n^e(K,x) &= \frac{\sqrt{n}(\widehat{g}_n^e(K,x) - g_0(x))}{V_n(K,x)^{1/2}} - \frac{\sqrt{n}(\widehat{g}_n(K,x) - g_0(x))}{V_n(K,x)^{1/2}} \\ &= t_n^e(K,x) + R_n^e(K,x) - R_n(K,x) \end{aligned}$$

where $t_n^e(K,x) = n^{-1/2}\sum_{i=1}^n (e_i - 1) f_{n,K,x}(\varepsilon_i, x_i)$, $R_n^e(K,x) = R_1^e(K,x) + R_2^e(K,x)$, $R_1^e(K,x)$, and $R_2^e(K,x)$ are defined the same as in Lemma 1 with the rescaled data $\{(\sqrt{e_i}\tilde{P}(K,x_i), \sqrt{e_i}\varepsilon_i\}_{i=1}^n$. Note that $\widehat{\beta}_K^e$ is the weighted least square estimator for the original data, and we can extend the uniform linearization results in Lemma 2 by replacing ζ_K with $\zeta_K^e = \zeta_K \log^{1/2} n$ and noting that

$E[e_i] = 1, E[e_i^2] = 1, \max_{1\le i\le n} |e_i| = o_p(\log n)$.

By applying Theorem 2.1 in Chernozhukov et al. (2016) to the weighted bootstrap process $t_n^e(K,x)$, there exists a random variable $\widetilde{Z}^e \overset{d|X^n}{=} \sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |Z_n(K,x)|$ such that, for every $\gamma \in (0,1)$,

$$P(|\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |t_n^e(K,x)| - \widetilde{Z}^e| > C_3\delta_{2n}) \le C_4(\gamma + n^{-1}) \tag{A.16}$$

where C_3, C_4 are positive constants that depend only on q,

$$\delta_{2n} = \gamma^{-1/q} n^{-1/2+1/q} \max_K \zeta_K \log^2 n + \gamma^{-1/3} n^{-1/6} (\max_K \zeta_K)^{1/3} \log n,$$

and $\overset{d|X^n}{=}$ denotes that the two random variables have the same conditional distribution given X^n.

Further,

$$\big|\sup_{K,x} |\widehat{T}_n^e(K,x)| - \sup_{K,x} |t_n^e(K,x)|\big| \le \sup_{K,x} |\widehat{T}_n^e(K,x) - T_n^e(K,x)| + \sup_{K,x} |T_n^e(K,x) - t_n^e(K,x)| = o_p(a_n)$$

by using $E\big[\sup_{K,x} |t_n^e(K,x)|\big] \le \max_{1\le i\le n} |e_i| E\big[\sup_{K,x} |t_n(K,x)|\big] \lesssim_P \log^{3/2} n$, Assumption 4.1(iv), and $|R_n^e(K,x)| = o_p(a_n), |R_n(K,x)| = o_p(a_n)$ uniformly in $(K,x) \in \mathcal{K}_n \times \mathcal{X}$ under the rate conditions in Assumption 4.1(ii) with $a_n = 1/(\log n)^{1/2}$. Then, there exists some sequence of positive constant δ_{3n}, δ_{4n} such that $\delta_{3n} = o(1), \delta_{4n} = o(1)$,

$$P(\big|\sup_{K,x} |\widehat{T}_n^e(K,x)| - \sup_{K,x} |t_n^e(K,x)|\big| > a_n\delta_{3n}) \le \delta_{4n}. \tag{A.17}$$

Combining (A.16) and (A.17) gives

$$P(|\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n^e(K,x)| - \widetilde{Z}^e| > a_n\delta_{3n} + C_3\delta_{2n}) \le C_4(\gamma + n^{-1}) + \delta_{4n} \tag{A.18}$$

By the Markov's inequality, the following is deduced from (A.18), for every $\nu \in (0,1)$,

$$P(|\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n^e(K,x)| - \widetilde{Z}^e| > a_n\delta_{3n} + C_3\delta_{2n}|X^n) \le \nu^{-1}(C_4(\gamma + n^{-1}) + \delta_{4n}) \tag{A.19}$$

with probability at least $1-\nu$. Similar derivation as in Theorem 3.1 using Lemma A.14 gives

$$\sup_{u\in\mathbb{R}} \big|P(\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n^e(K,x)| \le u|X^n) - P(\widetilde{Z} \le u)\big| = (a_n\delta_{3n} + C_3\delta_{2n})\sqrt{\log n} + \nu^{-1}(C_4(\gamma + n^{-1}) + \delta_{4n}) \tag{A.20}$$

with probability at least $1-\nu$ where we use $\widetilde{Z}^e \overset{d|X^n}{=} \widetilde{Z}$ and $E[\sup_{K,x} |Z_n(K,x)|] \lesssim \sqrt{\log n}$. By taking $\gamma = (\log n)^{-1/2}$ and $\nu = \nu_n \to 0$ sufficiently slower than $(\log n)^{-1/2} \vee \delta_{4n}$, and using $\delta_{2n} = o(a_n)$, the rate conditions imposed in the theorem, (A.20) is $o_p(1)$. Combining this with (A.15),

$$\sup_{u\in\mathbb{R}} \big|P(\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n^e(K,x)| \le u|X) - P(\sup_{(K,x)\in\mathcal{K}_n\times\mathcal{X}} |\widehat{T}_n(K,x)| \le u)\big| = o_p(1). \tag{A.21}$$

Then, the coverage result (4.7) follows. The second part of the theorem, (4.8), can be similarly derived as in the proof of Theorem 3.1 and this completes the proof. ■

A.2.3 Proof of Theorem 5.1

Proof. Conditional on $X = [x_1, \cdots, x_n]'$, the following decomposition holds for any sequence $K \in \mathcal{K}_n$:

$$\sqrt{n}(\widehat{\theta}_n(K) - \theta_0) = \widehat{\Gamma}_n(K)^{-1} S_n(K), \quad \widehat{\Gamma}_n(K) = \frac{1}{n}(W' M_K W), \quad S_n(K) = \frac{1}{\sqrt{n}} W' M_K (g + \varepsilon)$$

where $g = [g_1, \cdots, g_n]', g_i = g_0(x_i), g_w = [g_{w1}, \cdots, g_{wn}]', g_{wi} = g_{w0}(x_i) = E[w_i|x_i]$, and $v = [v_1, \cdots, v_n]$. All remaining proofs contain conditional expectations (conditioning on X) and hold almost surely (a.s.). Under Assumption 5.2,

$$\widehat{\Gamma}_n(K) = \Gamma_n(K) + o_p(1), \quad \Gamma_n(K) = \frac{1}{n} \sum_{i=1}^{n} M_{K,ii} E[v_i^2|x_i]$$

by Lemma 1 of Cattaneo, Jansson, and Newey (2018a). Moreover,

$$S_n(K) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} M_{K,ii} v_i \varepsilon_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1, j<i}^{n} P_{K,ij}(v_i \varepsilon_j + v_j \varepsilon_i) + o_p(1)$$

since $M_{K,ij} = -P_{K,ij}$ for $j < i$, $\frac{1}{\sqrt{n}} g_w' M_K g = O_p(\sqrt{n} \overline{K}^{-\gamma_g - \gamma_{g_w}}) = o_p(1)$, $\frac{1}{\sqrt{n}}(v' M_K g + g_w' M_K \varepsilon) = O_p(\overline{K}^{-\gamma_g} + \overline{K}^{-\gamma_{g_w}}) = o_p(1)$ by Lemma 2 of Cattaneo, Jansson and Newey (2018a) under Assumption 5.2. Then, the following holds:

$$T_n(K, \theta_0) = \sqrt{n} V_n(K)^{-1/2} (\widehat{\theta}_n(K) - \theta_0) = V_n(K)^{-1/2} \Gamma_n(K)^{-1} \frac{1}{\sqrt{n}} v' M_K \varepsilon + o_p(1) \xrightarrow{d} N(0, 1)$$

by Theorem 1 of Cattaneo, Jansson and Newey (2018a).

For simplicity, here we only show the joint convergence of bivariate t-statistics, but the proof can be easily extended to the multivariate case. For any $K_1 < K_2$ in $\mathcal{K}_n$, we show

$$Y_n = \Xi^{-1/2}(\delta_1 T_n(K_1, \theta_0) + \delta_2 T_n(K_2, \theta_0)) \xrightarrow{d} N(0, 1), \quad \forall (\delta_1, \delta_2) \in \mathbb{R}^2 \tag{A.22}$$

where $\Xi = \delta_1^2 + \delta_2^2 + 2\delta_1 \delta_2 v_{12}, v_{12} = \lim_{n \to \infty} V_n(K_1)^{-1/2} \Gamma_n(K_1)^{-1} \Omega_n(K_1, K_2) \Gamma_n(K_2)^{-1} V_n(K_2)^{-1/2}$.

Define $Y_n = Y_{1,n} + Y_{2,n}, Y_{1,n}$ and $Y_{2,n}$ as follows

$$Y_{1,n} = \omega_{1,1n} + \sum_{i=2}^{n} y_{1,in}, y_{1,in} = \omega_{1,in} + \bar{y}_{1,in}, \quad Y_{2,n} = \omega_{2,1n} + \sum_{i=2}^{n} y_{2,in}, y_{2,in} = \omega_{2,in} + \bar{y}_{2,in},$$

where $\omega_{1,in} = b_{1,n} W_{1,in}$, $b_{1,n} = \delta_1 \Xi^{-1/2} V_n(K_1)^{-1/2} \Gamma_n(K_1)^{-1}$, $W_{1,in} = v_i M_{K_1,ii} \varepsilon_i / \sqrt{n}$, $\bar{y}_{1,in} = \sum_{j<i}(u_{1,j} P_{K_1,ij} \varepsilon_i + u_{1,i} P_{K_1,ij} \varepsilon_j)/\sqrt{K_1}, u_{1,i} = c_{1,n} v_i, c_{1,n} = -\delta_1 \Xi^{-1/2} V_n(K_1)^{-1/2} \Gamma_n(K_1)^{-1} \sqrt{K_1/n}$

and $\omega_{2,in} = b_{2,n}W_{2,in}, \bar{y}_{2,in}$ are similarly defined with $P_{K_2}, V_n(K_2), \Gamma_n(K_2)$ and K_2. Note that $||V_n(K_1)^{-1}|| \leq C$ and $||\Gamma_n(K_1)^{-1}|| \leq C$ a.s. for n large enough by Assumption 5.2, and it follows that $||b_{1,n}|| \leq C$. Also, $E[||\omega_{1,1n}||^4|X] \leq C\sum_{i=1}^n E[||W_{1,in}||^4|X] \to 0$ a.s. by Assumption 5.2(ii). Using the same arguments in the proof of Lemma A2 in Chao et al. (2012), we have $\omega_{1,1n} = o_p(1)$ and $\omega_{2,1n} = o_p(1)$ unconditionally, thus $Y_n = \sum_{i=2}^n y_{in} + o_p(1), y_{in} = y_{1,in} + y_{2,in}$.

Let $\mathcal{X}_i = (W_{1,in}, W_{2,in}, v_i, \varepsilon_i)'$ and define the σ-fields $F_{i,n} = \sigma(\mathcal{X}_1, ..., \mathcal{X}_i)$ for $i = 1, ..., n$. Then, conditional on X, $\{y_{in}, F_{i,n}, 1 \leq i \leq n, n \geq 2\}$ is a martingale difference array with $F_{i-1,n} \subseteq F_{i,n}$. We apply the martingale central limit theorem to show, conditional on X, $\sum_{i=2}^n y_{in} \xrightarrow{d} N(0,1)$ a.s. Note that $E[\omega_{1,in}\bar{y}_{1,jn}|X] = 0, E[\omega_{1,in}\bar{y}_{2,jn}|X] = 0, E[\omega_{2,in}\bar{y}_{1,jn}|X] = 0, E[\omega_{2,in}\bar{y}_{2,jn}|X] = 0$ for all i, j. Then similar to the proof of Lemma A2 in Chao et al. (2012),

$$\begin{aligned}
s_n^2(X) = E[(\sum_{i=2}^n y_{in})^2|X] &= \sum_{i=2}^n (E[\omega_{1,in}^2|X] + E[\bar{y}_{1,in}^2|X]) + \sum_{i=2}^n (E[\omega_{2,in}^2|X] + E[\bar{y}_{2,in}^2|X]) \\
&+ 2\sum_{i=2}^n (E[\omega_{1,in}\omega_{2,in}|X] + E[\bar{y}_{1,in}\bar{y}_{2,in}|X]) \\
&= \delta_1^2\Xi^{-1} + \delta_2^2\Xi^{-1} - E[\omega_{1,1n}^2|X] - E[\omega_{2,1n}^2|X] - 2E[\omega_{1,1n}\omega_{2,1n}|X] \\
&+ 2\delta_1\delta_2\Xi^{-1}V_n(K_1)^{-1/2}\Gamma_n(K_1)^{-1}\Omega_n(K_1, K_2)\Gamma_n(K_2)^{-1}V_n(K_2)^{-1/2} \to 1 \quad a.s.
\end{aligned}$$

Moreover, we have $\sum_{i=2}^n E[y_{in}^4|X] \lesssim \sum_{i=2}^n E[y_{1,in}^4|X] + \sum_{i=2}^n E[y_{2,in}^4|X] \xrightarrow{a.s.} 0$ as in the proof of Lemma A2 of Chao et al. (2012).

It remains to prove that for any $\delta > 0$, $P(\big|\sum_{i=2}^n E[y_{in}^2|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - s_n^2(X)\big| \geq \delta|X) \to 0$. Note that

$$\sum_{i=2}^n E[y_{in}^2|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - s_n^2(X)$$

$$= \sum_{i=2}^n E[y_{1,in}^2|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - \sum_{i=2}^n (E[\omega_{1,in}^2|X] + E[\bar{y}_{1,in}^2|X]) \tag{A.23}$$

$$+ \sum_{i=2}^n E[y_{2,in}^2|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - \sum_{i=2}^n (E[\omega_{2,in}^2|X] + E[\bar{y}_{2,in}^2|X]) \tag{A.24}$$

$$+ 2\Big(\sum_{i=2}^n (E[\omega_{1,in}\omega_{2,in}|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - E[\omega_{1,in}\omega_{2,in}|X]) \tag{A.25}$$

$$+ \sum_{i=2}^n E[\omega_{1,in}\bar{y}_{2,in} + \omega_{2,in}\bar{y}_{1,in}|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] + \sum_{i=2}^n (E[\bar{y}_{1,in}\bar{y}_{2,in}|\mathcal{X}_1, ..., \mathcal{X}_{i-1}, X] - E[\bar{y}_{1,in}\bar{y}_{2,in}|X])\Big). \tag{A.26}$$

(A.23) and (A.24) converge to 0 a.s. by the proof of Lemma A2 in Chao et al. (2012). Moreover, it is straightforward to verify that (A.25) and (A.26) converge to 0 a.s. since $P_{K_1,ij}P_{K_2,ij} \leq P_{K_1,ij}^2 \vee P_{K_2,ij}^2$, $K_1 \asymp K_2$ and by closely following the proof of Lemma A2 in Chao et al. (2012). Then we can apply the martingale central limit theorem and deduce $Y_n \xrightarrow{d} N(0,1)$ using similar

arguments to the proof of Lemma A2 in Chao et al. (2012). Coverage results (5.6) and (5.7) follow by the joint convergence of $\widehat{T}_n(K, \theta_0)$ with $\max_{K \in \mathcal{K}_n} |\frac{\widehat{V}_n(K)}{V_n(K)} - 1| = o_p(1), ||\widehat{\Sigma}_n - \Sigma_n|| = o_p(1)$ as $n, K \to \infty$ under the assumption imposed in Theorem 5.1 and the Slutzky theorem. This completes the proof. ■

Appendix B Figures and Tables

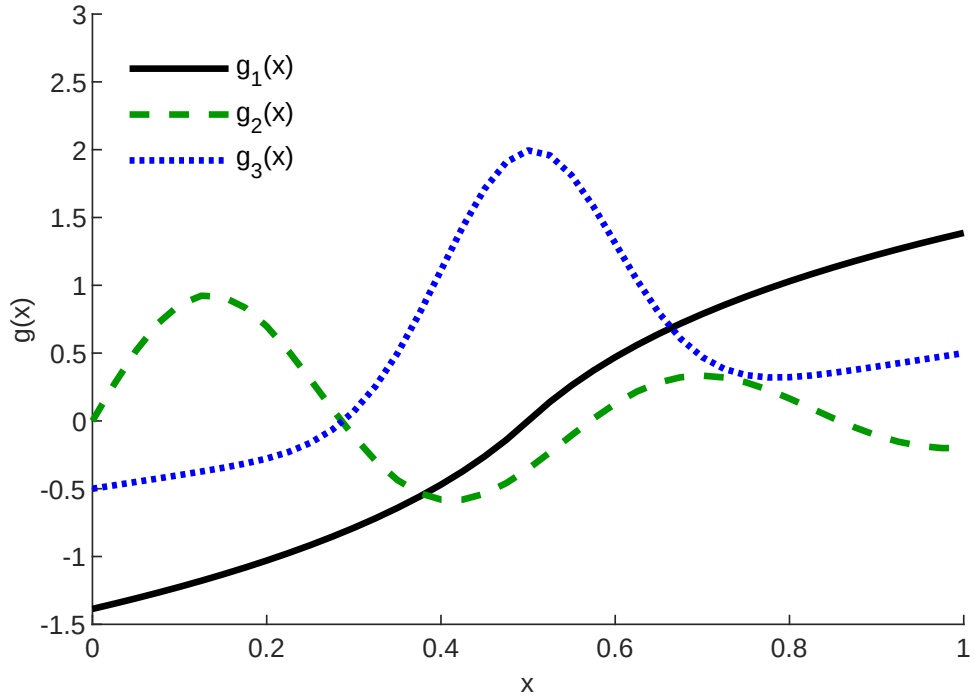

Figure 1: Different functions of $g(x)$ used in simulations (Section 6).
Solid lines (Black) are $g_1(x) = \ln(|6x-3|+1)sgn(x-1/2)$; Dashed lines (Green) are $g_2(x) = \sin(7\pi x/2)/[1+2x^2(sgn(x)+1)]$; Dotted lines (Blue) are $g_3(x) = x - 1/2 + 5\phi(10(x-1/2))$, where $\phi(\cdot)$ is the standard normal pdf.

Table 1: Coverage and Length of Nominal 95% CIs and CBs - Splines

	Pointwise								Uniform	
	$x = 0.2$		$x = 0.5$		$x = 0.8$		$x = 0.9$			
	COV	AL	COV	AL	COV	AL	COV	AL	COV	AL
Model 1: $g_1(x) = \ln(\lvert 6x-3\rvert + 1)sgn(x-1/2)$										
Standard	0.93	0.27	0.93	0.36	0.91	0.92	0.92	1.49	0.42	0.69
Robust ($\widehat{K}_{\texttt{cv}}$)	0.98	0.37	0.98	0.46	0.96	1.14	0.95	1.76	0.97	1.33
Robust ($\widehat{K}_{\texttt{cv+}}$)	0.98	0.51	0.98	0.49	0.98	1.51	0.97	2.08	0.98	1.42
Model 2: $g_2(x) = \sin(7\pi x/2)/[1 + 2x^2(sgn(x) + 1)]$										
Standard	0.80	0.28	0.93	0.36	0.91	0.92	0.92	1.49	0.27	0.69
Robust ($\widehat{K}_{\texttt{cv}}$)	0.93	0.37	0.97	0.46	0.96	1.14	0.95	1.76	0.96	1.33
Robust ($\widehat{K}_{\texttt{cv+}}$)	0.98	0.51	0.98	0.49	0.98	1.51	0.97	2.08	0.98	1.42
Model 3: $g_3(x) = x - 1/2 + 5\phi(10(x-1/2))$										
Standard	0.77	0.29	0.65	0.40	0.89	1.00	0.91	1.57	0.16	0.70
Robust ($\widehat{K}_{\texttt{cv}}$)	0.88	0.39	0.74	0.50	0.96	1.23	0.95	1.85	0.75	1.35
Robust ($\widehat{K}_{\texttt{cv+}}$)	0.98	0.52	0.92	0.53	0.98	1.52	0.97	2.06	0.97	1.44

Notes: "Pointwise" reports coverage (COV) and average length (AL) of (1) the standard 95% CI with $\widehat{K}_{\texttt{cv}} \in \mathcal{K}_n$; (2) robust CI with $\widehat{K}_{\texttt{cv}}$; (3) robust CI with $\widehat{K}_{\texttt{cv+}}$. "Uniform" reports analogous uniform inference results for confidence bands. $\widehat{K}_{\texttt{cv}}$ is selected to minimize leave-one-out cross-validation and $\widehat{K}_{\texttt{cv+}} = \widehat{K}_{\texttt{cv}} + 2$. Using quadratic spline regressions with evenly placed knots.

Table 2: Nonparametric Wage Elasticity of Hours of Work Estimates in Blomquist and Newey (Table 1, 2002). Wage elasticity evaluated at the mean net wage rates, virtual income, and level of hours.

Additional Terms[1]	CV^2	$\widehat{E}_w$	$SE_{\widehat{E}_w}$	$CI_{\widehat{E}_w}(K)$
$1, y_J, w_J$	0.00472	0.0372	0.0104	[0.0168, 0.0576]
$\Delta y \Delta w$	0.0313	0.0761	0.0128	[0.0510, 0.1012]
$\ell \Delta y$	0.0305	0.0760	0.0127	[0.0511, 0.1009]
y_J^2, w_J^2	0.0323	0.0763	0.0129	[0.0510, 0.1016]
$\Delta y^2, \Delta w^2$	0.0369	0.0543	0.0151	[0.0247, 0.0839]
$y_J w_J$	0.0364	0.0659	0.0197	[0.0273, 0.1045]
$\Delta y w$	0.0350	0.0628	0.0223	[0.0191, 0.1065]
$\ell^2 \Delta y$	0.0364	0.0636	0.0223	[0.0199, 0.1073]
y_J^3, w_J^3	0.0331	0.0845	0.0275	[0.0306, 0.1384]
$\ell \Delta y^2, \ell \Delta w^2, \ell \Delta y w$	0.0263	0.0775	0.0286	[0.0214, 0.1336]
$y_J^2 w_J, y_J w_J^2$	0.0252	0.0714	0.0289	[0.0148, 0.1280]
MLE estimates		0.123	0.0137	
critical values: $\widehat{c}_{1-\alpha}(x) = 2.503$,		$CI^{\texttt{sup}}_{\widehat{E}_w}(\widehat{K}_{\texttt{cv}}) = [0.0165, 0.0921]$[3]		
$CI^{\texttt{sup}}_{\widehat{E}_w}(\widehat{K}_{\texttt{cv+}}) = [0.0166, 0.1152]$,		$CI^{\texttt{sup}}_{\widehat{E}_w}(\widehat{K}_{\texttt{cv++}}) = [0.0070, 0.1186]$		

[1] y : non-labor income, w : marginal wage rates, ℓ: the end point of the segment in a piecewise linear budget set. $\ell^m \Delta y^p w^q$ denotes $\sum_j \ell_j^m (y_j^p w_j^q - y_{j+1}^p w_{j+1}^q)$.

[2] CV denotes the cross-validation criteria defined in Blomquist and Newey (2002, p.2464). $\widehat{K}_{\texttt{cv}} = K_5$, the 5th smallest model, is chosen by the cross-validation, and let $\widehat{K}_{\texttt{cv+}} = K_6$, $\widehat{K}_{\texttt{cv++}} = K_7$.

[3] $CI^{\texttt{sup}}_{\widehat{E}_w}(K) = \widehat{E}_w(K) \pm \widehat{c}_{1-\alpha}(x) SE_{\widehat{E}_w}(K)$, $CI_{\widehat{E}_w}(K) = \widehat{E}_w(K) \pm z_{1-\alpha/2} SE_{\widehat{E}_w}(K)$.

Figure 2: Nonparametric Wage Elasticity of Hours of Work Estimates in Blomquist and Newey (Table 1, 2002).

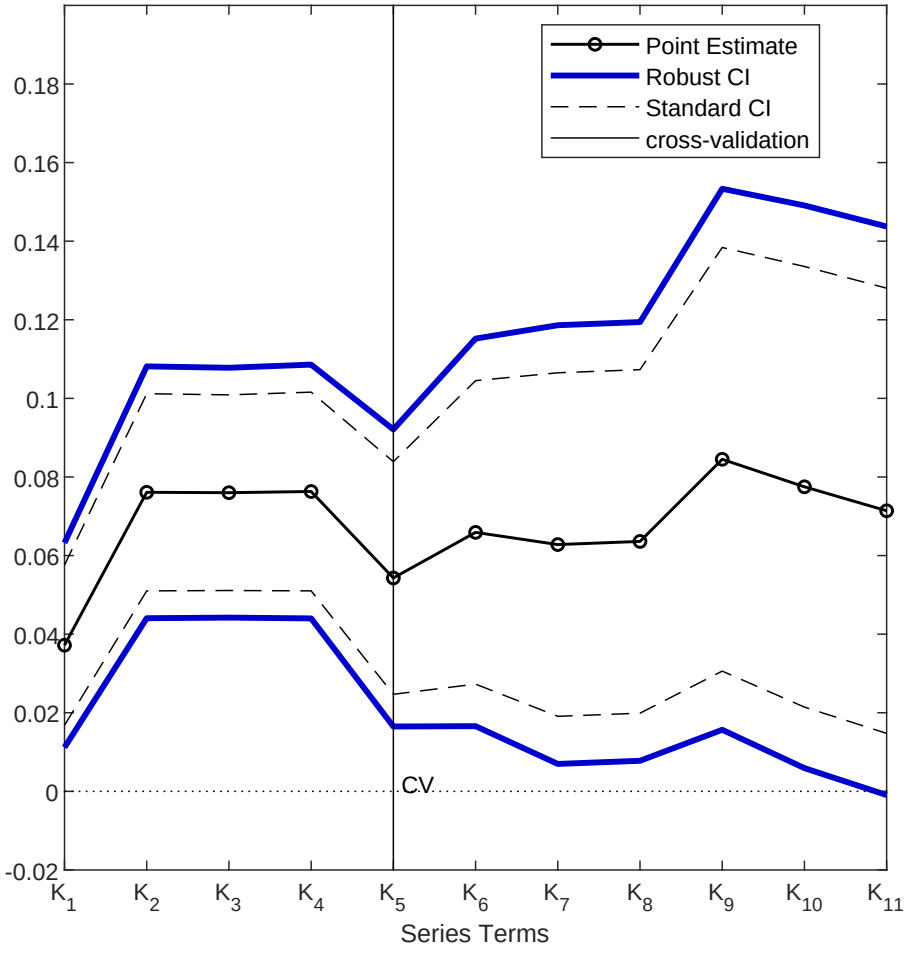

Figure 2 plots the wage elasticity estimates of the expected labor supply same as in Table 2, with standard pointwise 95% CIs as well as uniform (in $K \in \mathcal{K}_n$) CIs constructed with the critical value $\widehat{c}_{1-\alpha}(x)$.